FRONTIERSPATIAL: Holistic Evaluation of Spatial Reasoning in Multimodal Large Language Models

Wahid Faisal [*1 2] **Azmine Toushik Wasi** [*1 2] **Mohammed Eunus Ali** [3] **Md Rizwan Parvez** [4]

Abstract

Spatial reasoning is a core capability for intelligent systems, yet evaluation of multimodal large language models (MLLMs) remains fragmented across narrowly focused benchmarks targeting isolated skills. To address this limitation, we introduce FRONTIERSPATIAL, the first unified large-scale benchmark and evaluation harness for holistic spatial reasoning in multimodal models. FRONTIERSPATIAL consolidates and extends 23 datasets into 7,838 questions and 11,825 images, organized within a four-level hierarchy spanning 7 categories, 19 sub-categories, and 65 fine-grained tasks, including Physical, Geometric, Knowledge-Grounded, Relational/Causal, Temporal, Quantitative, and Perceptual Reasoning. Beyond the benchmark, we contribute: (1) an automated curation pipeline that distills over 60,000 raw instances into a balanced, high-quality benchmark; (2) an automated evaluation harness for reproducible assessment of 16 state-of-the-art MLLMs; and (3) an automated error-analysis framework that generates structured diagnostics across the task hierarchy. Experiments reveal large performance disparities across spatial reasoning categories and systematic weaknesses that remain hidden when benchmarks are studied independently. FRONTIERSPATIAL provides a rigorous and extensible foundation for holistic multimodal spatial reasoning research.

1. Introduction

Spatial reasoning is a cornerstone of human cognition, encompassing the ability to interpret geometric relations, infer depth and occlusion, track object motion, navigate environments, and manipulate mental representations of physical space. These capabilities underpin a broad spectrum of real-world applications, from robotics, autonomous driving, and augmented reality to geospatial analytics, medical imaging, and education (Huang et al., 2024; Qi et al., 2024; 2025). For artificial agents operating in dynamic environments, reliable spatial competence demands integration of visual cues, structural scene understanding, and multi-step inference across perspectives that are visually noisy and geometrically complex (Bar-Anan et al., 2006; Cai et al., 2025b; Trope & Liberman, 2010; Ramalho et al., 2018). Recent vision–language models (VLMs) have made remarkable strides in multimodal representation and language grounding, yet their spatial judgments remain fragile, inconsistent, and poorly understood under realistic deployment conditions (Chen et al., 2025; Kosoy et al., 2025; Pothiraj et al., 2025; Wasi et al., 2026).

A key obstacle to progress is the lack of a unified, comprehensive evaluation framework. The landscape of spatial reasoning benchmarks is highly fragmented, each existing suite targets a narrow slice of the problem: binary positional relations (Kamath et al., 2023; Du et al., 2024), low-resolution depth estimation (Cheng et al., 2024a), schematic navigation (Song et al., 2024), or domain-specific perception (Hu et al., 2023; Wang et al., 2023), often in synthetic or simplified visual settings that reduce perceptual demand and homogenize object appearances (Jia et al., 2025; Szymanska et al., 2024; Shiri et al., 2024; Cai et al., 2024; Yang et al., 2024; Qi et al., 2025). These controlled setups mask brittleness under distribution shift, produce apparent performance saturation, and leave critical challenges, robust occlusion inference, scale-consistent judgments across viewpoints, planning under partial observability, and causal spatial understanding, chronically under-sampled (Kosoy et al., 2025; Pothiraj et al., 2025; Chen et al., 2025). Consequently, models that score highly on established benchmarks often fail on tasks requiring integrated, multi-step spatial inference, and no single benchmark provides actionable signals across the full breadth of spatial cognition (**?**Cai et al., 2025a; Jiang et al., 2024).

As illustrated in Table 4, existing benchmarks cover only

[*]Equal contribution [1]Computational Intelligence and Operations Laboratory (CIOL) [2]Shahjalal University of Science and Technology, Bangladesh [3]Monash University, Australia [4]Qatar Computing Research Institute (QCRI), Qatar. Correspondence to: Azmine Toushik Wasi <azmine32@student.sust.edu>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

a narrow subset of the 22 spatial reasoning dimensions we identify. Benchmarks such as SPATIAL457 (Wang et al., 2025b), MMSI-BENCH (Yang et al., 2025), and SPATIALAB (Wasi et al., 2026) address several spatial sub-skills but remain confined to 4–6 dimensions each. Even broad multimodal benchmarks such as MMMU (Yue et al., 2023) and MM-VET (Yu et al., 2023) provide only incidental coverage of spatial tasks. No existing benchmark achieves simultaneous coverage across all core axes of spatial reasoning: physical interaction, temporal dynamics, geometric and 3D structure, logical and causal inference, quantitative comparison, knowledge-grounded understanding, and perceptual recognition. This fragmentation means that the community currently lacks a reliable, holistic picture of where VLMs stand on spatial intelligence.

To close this gap, we introduce FRONTIERSPATIAL, a holistic benchmark and evaluation framework for spatial reasoning in multimodal language models, inspired by HAL (Kapoor et al., 2026). Drawing from 23 existing datasets and applying an automated curation pipeline, FRONTIERSPATIAL assembles **7,838 questions paired with 11,825 images** under a principled four-level task hierarchy comprising **7 categories**, **19 sub-categories**, and **65 fine-grained tasks**. The taxonomy spans Physical Reasoning, Positional and Geometric Reasoning, Knowledge-Grounded Reasoning, Relational/Logical and Causal Reasoning, Temporal and Dynamic Reasoning, Quantitative and Comparative Reasoning, and Perceptual Understanding, together forming the first benchmark to achieve full coverage across all major dimensions of spatial cognition. Beyond the benchmark itself, we contribute an automated evaluation harness for reproducible assessment of 20–25 state-of-the-art VLMs, and an automated error analysis harness that produces structured, fine-grained diagnostics across the task hierarchy.

Our key contributions are as follows:

1. **Holistic Taxonomy and Benchmark.** We introduce a curated taxonomy of spatial reasoning organized into 7 categories, 19 sub-categories, and 65 tasks, developing the first unified evaluation covering all major axes of spatial cognition simultaneously.
2. **Automated Curation Pipeline and Curated Dataset.** We develop a scalable, automated pipeline that distills over 60,000 raw instances into a high-quality, balanced benchmark through taxonomic coverage, quality filtering, and deduplication. The resulting dataset comprises 7,838 QA pairs across 11,825 images sourced from 23 existing datasets, and the pipeline is designed to generalize to other reasoning domains beyond spatial understanding.
3. **Automated Evaluation Harness and Evaluation.** We provide a fully automated evaluation harness enabling reproducible, standardized assessment of 16 state-of-the-art open- and closed-source VLMs, including reasoning-focused and spatially specialized systems, with human baselines for absolute gap quantification.
4. **Automated Error Analysis Harness and Detailed Error Analysis.** We deliver a structured error analysis harness that generates fine-grained diagnostic reports across the task hierarchy, uncovering systematic failure modes and providing actionable insights into the spatial limitations of current VLMs.
5. **Agent-Based Hard Case Evaluation.** We evaluate the most challenging instances using two setups: (i) *SpatioXolver*, a multi-agent system, and (ii) a *Claude Code*-based agent. These configurations solve a subset of problems that remain unsolved by all evaluated frontier VLMs, highlighting the potential of agentic decomposition for difficult spatial reasoning tasks.

Our evaluation reveals substantial and consistent performance disparities across spatial reasoning categories, with models performing near chance on several sub-skills that humans handle reliably. In particular, the human–model gap is pronounced, with human performance at 91.22% compared to the best model at 45.20%, and several categories (e.g., Quantitative & Comparative reasoning) remaining below 17% across all systems. Even strong proprietary models such as GEMINI-3.1-FLASH-LITE and GPT-5-MINI fail to exceed mid-range performance on multiple reasoning axes, underscoring persistent structural limitations rather than isolated weaknesses. Failures concentrate in tasks requiring compositional multi-step inference, causal spatial understanding, and perspective-consistent 3D reasoning, limitations that are entirely invisible when models are evaluated on any single existing benchmark. FRONTIERSPATIAL establishes a rigorous, extensible foundation for holistic spatial reasoning research in the multimodal era.

2. Preliminaries: Spatial Reasoning in Large Multimodal Models

Formal Definition of Visual–Spatial Reasoning. Visual–spatial reasoning is the cognitive capacity to perceive, represent, and manipulate geometric and relational structures among entities in a scene, a faculty central to human intelligence and indispensable for embodied agents operating in dynamic, unstructured environments (Palmer, 1999; Chabris et al., 2006; Baddeley, 1998; Previc, 1998). Formally, let a visual scene be represented as a set of entities $\mathcal{O} = \{o_1, \ldots, o_n\}$, where each entity o_i carries a tuple of spatial attributes: $o_i = (\mathbf{p}_i,\ \mathbf{r}_i,\ s_i,\ \mathbf{v}_i)$ denoting position $\mathbf{p}_i \in \mathbb{R}^3$, orientation $\mathbf{r}_i \in SO(3)$, scale $s_i \in \mathbb{R}_+$, and velocity $\mathbf{v}_i \in \mathbb{R}^3$ (for dynamic scenes), all defined within a world coordinate frame $\mathcal{W}$. The observer introduces an additional egocentric frame $\mathcal{E}$, governed by a viewpoint transform $\mathbf{T} \in SE(3)$, such that agent-centric spatial predicates (e.g., *in-front-of*, *to-the-left*) must be resolved relative to $\mathcal{E}$ rather

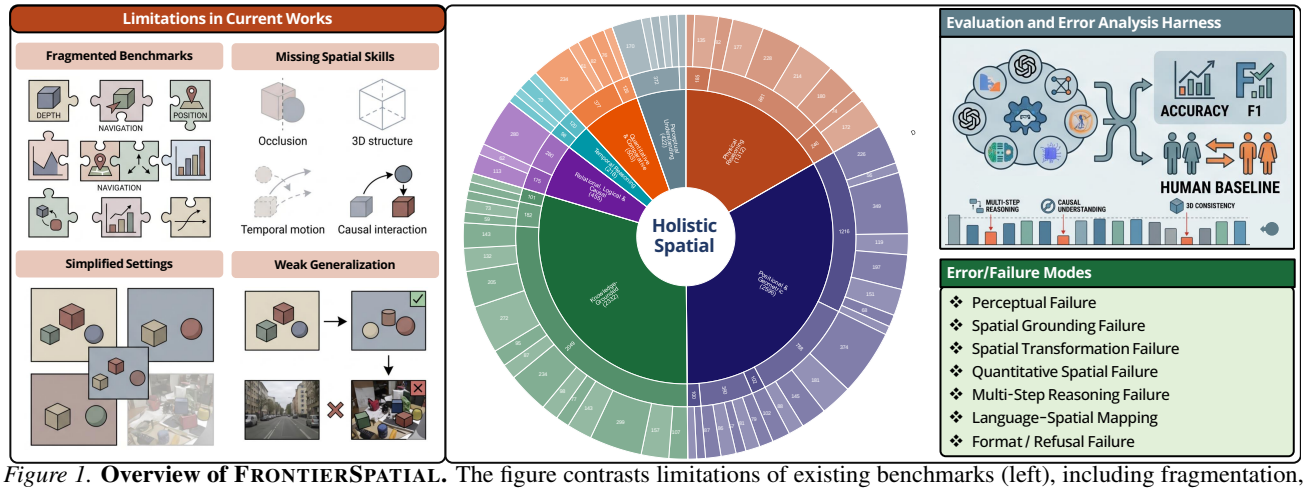


Figure 1. **Overview of FRONTIERSPATIAL.** The figure contrasts limitations of existing benchmarks (left), including fragmentation, simplified settings, missing spatial skills, and weak generalization, with the proposed unified taxonomy (center), spanning 7 categories, 19 sub-categories, and 65 tasks across diverse spatial dimensions. On the right, the evaluation and error analysis harness enables systematic assessment across models with standardized metrics and human baselines, while providing fine-grained diagnostic insights into failure modes such as perceptual errors, spatial grounding issues, and multi-step reasoning breakdowns.

than $\mathcal{W}$ (Trope & Liberman, 2010; Ramalho et al., 2018). A spatial relation between two entities is a predicate $r(o_i, o_j) \in \mathcal{R}$, where $\mathcal{R}$ is a structured relation vocabulary encompassing categories such as *positional* (e.g., *above*, *behind*), *geometric* (e.g., *parallel-to*, *intersects*), *physical* (e.g., *supports*, *occludes*), *causal* (e.g., *causes-motion*, *blocks*), and *temporal* (e.g., *precedes*, *co-occurs-with*). The full relational structure of a scene is then a typed graph: $\mathcal{G} = \big(\mathcal{O},\ \{(o_i, r_k, o_j) \mid o_i, o_j \in \mathcal{O},\ r_k \in \mathcal{R}\}\big)$

Visual–spatial reasoning in a vision–language model (VLM) seeks to learn a mapping: $f_\theta : (I,\ Q) \longmapsto A$ where $I \in \mathcal{I}$ is an image or image sequence, $Q \in \mathcal{Q}$ is a natural language query, and A is either a structured relational answer drawn from $\mathcal{G}$ or a free-form linguistic response. Learning f_θ is fundamentally difficult because it requires the simultaneous satisfaction of three constraints: *perceptual grounding* (extracting $\mathcal{O}$ and their attributes from I under noise, occlusion, and viewpoint change), *relational inference* (composing predicates in $\mathcal{R}$ across multiple entities), and *linguistic alignment* (mapping $\mathcal{G}$-structured answers to and from natural language under referential ambiguity) (Alayrac et al., 2022; OpenAI, 2023; Chen et al., 2025).

The Holistic Argument: Why Partial Evaluation is Insufficient. Prior work typically decomposes spatial reasoning into isolated sub-tasks, e.g., binary positional relations (Kamath et al., 2023; Du et al., 2024), metric depth estimation (Cheng et al., 2024b;a), object counting (Charoenpitaks et al., 2025), or navigation-centric grounding (Song et al., 2024). While such decomposition enables controlled evaluation, it misrepresents the nature of spatial cognition. Human spatial reasoning is inherently *integrative*: resolving queries such as *"which object would fall first if the shelf were removed?"* requires joint reasoning over perception (object detection), geometry (support relations), physics (gravity and contact), and causality, all under a consistent viewpoint (Gardner, 2011; Previc, 1998; Bar-Anan et al., 2006). Formally, we define a *holistic spatial query* as one whose resolution requires composition of predicates drawn from at least two distinct sub-vocabularies of $\mathcal{R}$. Let $\mathcal{R} = \mathcal{R}_{\text{pos}} \cup \mathcal{R}_{\text{geo}} \cup \mathcal{R}_{\text{phys}} \cup \mathcal{R}_{\text{caus}} \cup \mathcal{R}_{\text{temp}} \cup \mathcal{R}_{\text{quant}} \cup \mathcal{R}_{\text{percep}}$ denote the partition of relations corresponding to our seven taxonomy categories (§3). A benchmark $\mathcal{B}$ is *holistic* if (i) each sub-vocabulary is sufficiently represented, i.e., $\forall k \in \{1, \ldots, 7\},\ |\{q \in \mathcal{B} \mid \mathcal{R}_k \text{ is required}\}| \geq \tau$, for some threshold $\tau > 0$, and (ii) a non-trivial fraction of queries require cross-category composition. Under this definition, existing benchmarks are non-holistic: coverage is sparse and cross-category composition is rare (Table 4).

The consequence is practical. Strong performance on a benchmark $\mathcal{B}_k$ covering only $\mathcal{R}_k$ does not imply competence on $\mathcal{R}_{k'}$ for $k' \neq k$, nor on queries requiring compositional reasoning across $\mathcal{R}_k \times \mathcal{R}_{k'}$ (Kosoy et al., 2025; Pothiraj et al., 2025). Since real-world applications, e.g., robotics, autonomous driving, and augmented reality, require such composition (Huang et al., 2024; Collaboration et al., 2023; Qi et al., 2025), narrow benchmarks are poor predictors of deployment readiness. This gap motivates FRONTIERSPATIAL.

A Taxonomy of Spatial Reasoning. To operationalize holistic evaluation, we organize spatial reasoning into a four-level hierarchy. At the top level, we identify seven *categories* corresponding to the sub-vocabularies of $\mathcal{R}$ defined above, each grounded in a distinct cognitive faculty documented in the spatial cognition literature (Baddeley, 1998; Previc, 1998; Palmer, 1999) (Table 6):

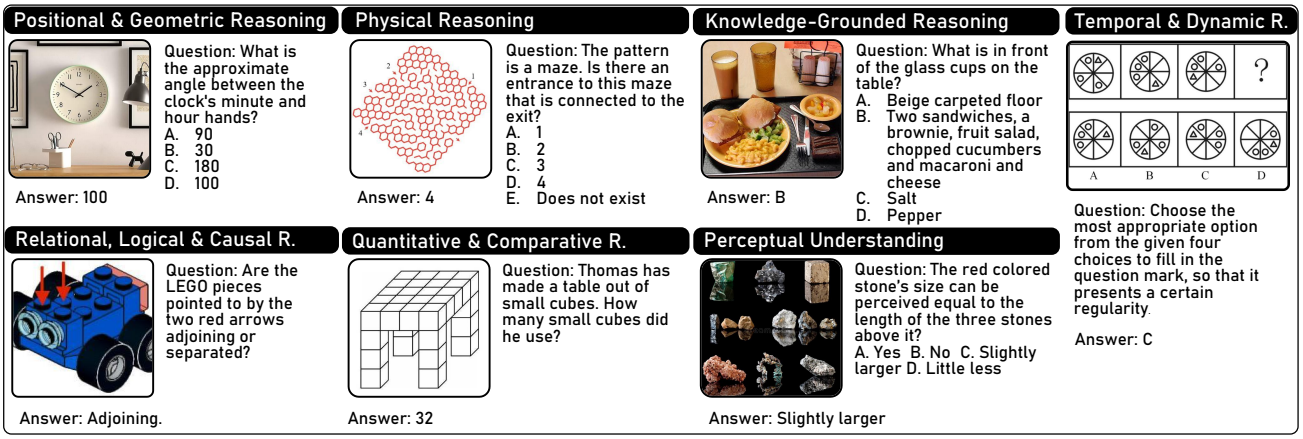


Figure 2. Representative examples from FRONTIERSPATIAL.

1. **Physical Reasoning** ($\mathcal{R}_{\text{phys}}$, $n = 1{,}312$): Inference about forces, stability, support, contact, and material properties. Requires integrating visual appearance with intuitive physics (Bar-Anan et al., 2006).
2. **Positional & Geometric Reasoning** ($\mathcal{R}_{\text{pos}} \cup \mathcal{R}_{\text{geo}}$, $n = 2{,}596$): Resolution of spatial predicates such as *left-of*, *above*, *parallel*, and metric distance, requiring consistent frame-of-reference management under viewpoint change (Trope & Liberman, 2010; Ramalho et al., 2018).
3. **Knowledge-Grounded Reasoning** ($\mathcal{R}_{\text{know}}$, $n = 2{,}332$): Spatial inference that must be conditioned on domain knowledge, geographic, architectural, or functional, linking perceptual cues to learned world models (Yue et al., 2023; Wang et al., 2025a).
4. **Relational, Logical & Causal Reasoning** ($\mathcal{R}_{\text{caus}}$, $n = 455$): Multi-step inference over entity relations, including counterfactual and causal queries (e.g., *"what would move if X were removed?"*) (Ma et al., 2026b; Jiang et al., 2024).
5. **Temporal & Dynamic Reasoning** ($\mathcal{R}_{\text{temp}}$, $n = 218$): Tracking spatial configurations across time, inferring motion trajectories, and predicting future states (Charoenpitaks et al., 2025; Fu et al., 2025).
6. **Quantitative & Comparative Reasoning** ($\mathcal{R}_{\text{quant}}$, $n = 503$): Metric and ordinal judgments over spatial properties, distance, count, and angle, requiring calibrated visual measurement (Cheng et al., 2024b; Wang et al., 2024).
7. **Perceptual Understanding** ($\mathcal{R}_{\text{percep}}$, $n = 422$): Low-level spatial perception, object localization, scene-level layout recognition, and textual spatial cues, forming the perceptual substrate upon which higher-order reasoning operates (Yu et al., 2023; Hu et al., 2023).

Each category is further decomposed into *sub-categories* (19 total) and *tasks* (65 total), forming a four-level hierarchy (Table 6): Category $\rightarrow$ Sub-category $\rightarrow$ Task $\rightarrow$ Instance. This granularity enables targeted diagnostics: a model may achieve strong aggregate accuracy while exhibiting categorical failure in, say, temporal reasoning or occlusion-conditioned physical inference, a pattern that flat benchmarks cannot expose.

Evaluation Framework. Let $\mathcal{B} = \{(I_i, Q_i, A_i^*)\}_{i=1}^N$ denote the benchmark, where A_i^* is the ground-truth answer. The model generates a free-form response $\hat{A}_i = f_\theta(I_i, Q_i)$, evaluated against A_i^* using an automated judge (§3.2,D). This format demands genuine generative spatial reasoning and compositional description, exposing weaknesses masked by multiple-choice formats (Yu et al., 2023; Wasi et al., 2026; Liu et al., 2025). Together, these protocols reflect the dual demands of real-world spatial reasoning, recognition under constraint and generation under ambiguity, and form the basis of the automated evaluation and error analysis harnesses described in §5.2 and E.

3. FRONTIERSPATIAL

FRONTIERSPATIAL is a large-scale, holistic benchmark for evaluating spatial reasoning in vision–language models (VLMs), constructed by consolidating and curating instances from 23 existing datasets through a four-stage automated pipeline (§C). The benchmark comprises **7,838 visual question–answer pairs** paired with **11,825 images**, organized under a principled four-level task hierarchy of **7 categories**, **19 sub-categories**, and **65 fine-grained tasks** (§B, Table 6). Unlike prior spatial benchmarks that target narrow slices of the problem space (Kamath et al., 2023; Du et al., 2024; Cheng et al., 2024b; Song et al., 2024), FRONTIERSPATIAL achieves simultaneous coverage across all major dimensions of spatial cognition, from physical interaction and 3D geometry to temporal dynamics, causal inference, and knowledge-grounded spatial understanding. Both multiple-choice and open-ended evaluation formats are supported, enabling assessment of discriminative accuracy alongside generative spatial reasoning. Table 4 illustrates

that FRONTIERSPATIAL is the only benchmark achieving full coverage across all 22 spatial reasoning dimensions identified in our taxonomy. Some examples are given in Figure 2, more examples with model responses and evaluations are available in §K.

3.1. Data Curation Pipeline

The benchmark was constructed via a four-stage automated curation pipeline designed to be principled, reproducible, and transferable to other reasoning domains, as illustrated in Figure 3. Starting from a raw pool of more than 60,000 candidate instances aggregated from 23 source datasets, the pipeline applies increasingly expensive quality criteria at each stage, reducing the pool to the final 7,838 high-quality instances. A summary of stage-wise reductions is provided in Table 7; full technical details are given in §C.

Stage 1: Structural heuristics (60K+ $\rightarrow$ 24,830). Five deterministic, LLM-free filters are applied in sequence: metadata integrity validation (ensuring all image paths resolve and answer fields are non-empty), publication recency filtering (removing pre-2020 annotation artifacts), question length filtering ($[10, 200]$ tokens), SHA-256-based exact-text deduplication across source boundaries, and taxonomy remapping with non-spatial pruning. The last step operationalizes the FRONTIERSPATIAL taxonomy as a quality gate, any instance that cannot be mapped to the 65-task hierarchy is removed. Stage 1 eliminates $\sim$50% of the raw pool at zero LLM cost.

Stage 2: Vision-space deduplication (24,830 $\rightarrow$ 12,914). All images are encoded with a CLIP ViT-L/14 encoder and indexed in a FAISS flat index. Near-duplicate pairs (cosine similarity > 0.97) are clustered and deduplicated, retaining the instance with the highest text-quality signal per cluster. Subsequently, K-means clustering ($K = 25$) over image embeddings identifies visual domain clusters; over-represented domains ($> 15\%$ of pool) are stratified-downsampled to ensure visual balance. This stage reveals that $\sim$37% of Stage 1 survivors contain near-duplicate visual content across source boundaries, a cross-dataset redundancy undetectable by text-level filtering alone.

Stage 3: Text-only LLM scoring (12,914 $\rightarrow$ 9,812). Each surviving instance is scored on four criteria via a single structured LLM call operating on text only: **R1** spatial relevance (threshold ≥ 3), **R2** cognitive depth (threshold ≥ 2), **R3** answerability from visual information alone (threshold ≥ 3), and **R4** language clarity (threshold ≥ 3). All four thresholds must be simultaneously satisfied. Operating on text only reduces cost to approximately one-tenth of full multimodal evaluation, while R1 and R3 together eliminate instances with language shortcuts or knowledge dependencies that circumvent genuine visual spatial reasoning.

Stage 4: Multimodal LLM final pass (9,812 $\rightarrow$ 7,838). Three multimodal criteria are applied to the surviving pool. **R5** (visual necessity) presents the model with question and choices *without* the image: instances answered correctly from text alone are discarded as harbouring language bias (Goyal et al., 2017; Kosoy et al., 2025). **R6** (image-question coherence) verifies that the referenced spatial scene and objects are present in the image. **R7** (spatial signal quality) scores the richness of spatial structure in the image on a 1–4 scale; instances scoring ≤ 2 are removed, and R7 serves as the final ranking signal within taxonomy cells when quota constraints apply.

3.1.1. BENCHMARK STATISTICS AND COVERAGE

Table 1 reports instance distribution across categories. The taxonomy is intentionally unbalanced at the category level, reflecting the natural distribution of spatial reasoning sub-skills across source datasets, an enforced uniform distribution would artificially inflate under-resourced categories. At the sub-category and task levels, minimum coverage thresholds are maintained to ensure every fine-grained skill is evaluable. The benchmark draws from 23 source datasets spanning indoor and outdoor scenes, remote sensing, biological imagery, video frames, and synthetic environments, ensuring visual domain diversity validated by the Stage 2 clustering procedure. Full dataset composition, task and category descriptions, per-task instance counts, and source attribution are provided in §B.

Table 1. Distribution of instances across the seven top-level categories of FRONTIERSPATIAL.

Category	# Questions	# Sub-cats	# Tasks
Physical Reasoning (Phys.)	1,312	3	9
Positional & Geometric Reasoning (Pos.Geo.)	2,596	4	18
Knowledge-Grounded Reasoning (Knowledge)	2,332	4	14
Relational, Logical & Causal Reasoning (Re.Lo.Ca.)	455	3	10
Temporal & Dynamic Reasoning (Temporal)	218	2	6
Quantitative & Comparative Reasoning (Quant.)	503	2	5
Perceptual Understanding (Percept.)	422	1	3
Total	**7,838**	**19**	**65**

3.2. Automated Evaluation Harness

To enable reproducible, large-scale scoring across different models and thousands of spatial reasoning tasks, we developed HARNESSEVALUATOR, a three-stage automated pipeline for structured signal extraction, error taxonomy induction, and cross-model statistical synthesis. At first, each model response is passed to an LLM judge under a zero-shot structured prompt that produces four binary scores per instance: *accuracy*, *spatial elements*, *spatial relations*, and *spatial reasoning*. The six-dimensional decomposition is intentional: overall accuracy conflates distinct competencies, whereas the sub-scores allow failures to be localised to the perception–grounding–reasoning pipeline rather than treated as a monolithic binary outcome. More details are available in §D). The pipeline is fully resumable, already-scored rows are detected and skipped on re-runs, making it practical to extend the evaluation to new models or benchmark items without reprocessing the full dataset. All code and instructions required to reproduce the experiments are available via the Code URL provided in the submission form.

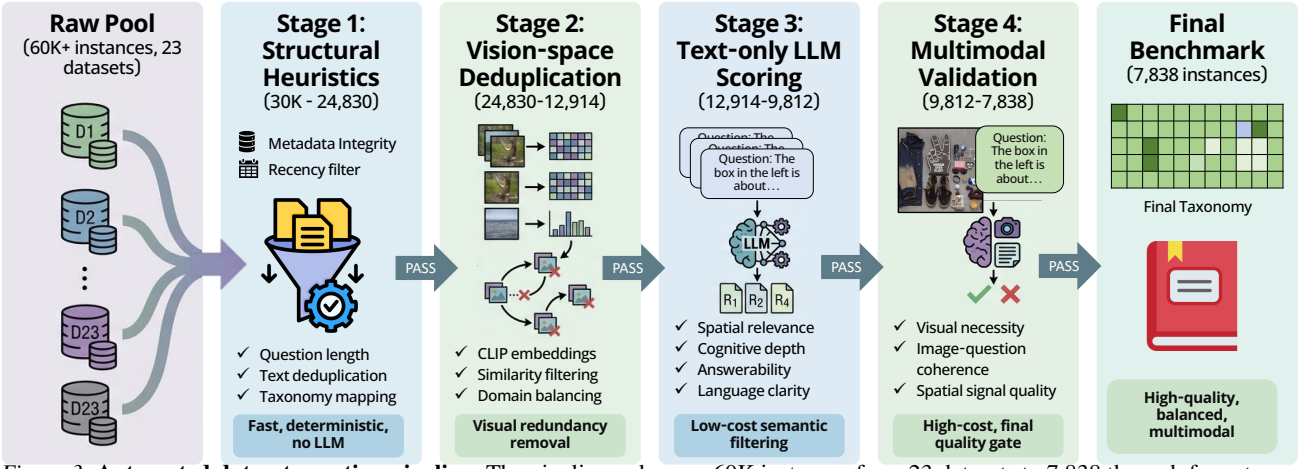


Figure 3. **Automated dataset curation pipeline.** The pipeline reduces ∼60K instances from 23 datasets to 7,838 through four stages: structural filtering, vision-space deduplication, text-only LLM scoring, and multimodal validation. Each stage removes low-quality or redundant data while increasing computational cost, resulting in a balanced and high-quality benchmark.

3.3. Error Analysis Pipeline

Standard evaluation in visual reasoning benchmarks reports aggregate or category-level accuracy, capturing *what* models get wrong but not *why*. This obscures heterogeneous failure mechanisms and limits diagnostic insight. To address this, we introduce an *Automated Error Analysis Pipeline* for fine-grained, reproducible analysis of spatial reasoning failures.

The pipeline comprises three stages: **(1) structured signal extraction**, **(2) inductive error taxonomy induction**, and **(3) cross-model diagnostic synthesis**, transforming raw predictions into interpretable, statistically grounded insights.

Stage 1: Structured Signal Extraction. Each prediction is mapped to a multi-attribute representation capturing correctness and failure characteristics, including answer validity, and reasoning coherence. This separates qualitatively distinct failure modes (e.g., formatting, reasoning, perceptual errors) and partitions incorrect predictions into disjoint error pools for homogeneous analysis.

Stage 2: Error Taxonomy Induction. We adopt a bottom-up approach, deriving error types directly from failures. Each incorrect instance (question, gold answer, prediction) is embedded into a failure-aware semantic space capturing *patterns of mistakes*. Density-based clustering (HDBSCAN) identifies recurring failure modes, which are automatically labeled via an LLM using representative examples. These clusters are then mapped to a fixed reference taxonomy (e.g., perceptual, grounding, transformation, multi-step reasoning) for consistent cross-model comparison.

Stage 3: Cross-Model Diagnostic Synthesis. Error labels are aggregated into per-model *error profiles* and *skill vectors*, capturing both failure patterns and spatial competencies. We apply paired tests (McNemar), bootstrap confidence intervals, and effect size estimation with multiple-comparison correction for rigorous comparison. An LLM-based synthesis module generates structured, evidence-backed diagnostic reports.

This pipeline yields insights beyond standard evaluation: error types transcend benchmark categories, models with similar accuracy exhibit distinct error profiles, and key bottlenecks (e.g., spatial transformation, quantitative reasoning) persist across models with large human–model gaps. The induced taxonomy is also stable across models, indicating it captures intrinsic structure in spatial reasoning. Further details are available in §E. The code and instructions for reproducing the experiments are provided via the Code URL in the submission form.

4. Experiments and Evaluation

We evaluate a diverse set of 16 open- and closed-source models on FRONTIERSPATIAL.

Models. We categorize the evaluated models into two groups. **Proprietary models** include Claude 4.5 Haiku (Anthropic, 2025), Gemini-2.5-Flash (Team, 2025), Gemini-3.1-Flash-Lite (Google DeepMind, 2025), GPT-4o-mini (OpenAI, 2024), GPT-5-mini (OpenAI, 2026), and o4-mini (OpenAI, 2025). **Open-source models** comprise Gemma-3-4B-it (Gemma Team, Google DeepMind, 2025), Qwen3-VL-8B-Instruct (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026a), Llama-3.2-11B-Vision-Instruct (Grattafiori et al., 2024), Gemma-4-31B-it (Gemma Team, Google DeepMind, 2026), Aya-Vision-32B (Dash et al., 2025), Qwen3.6-35B-A3B (Qwen Team, 2026b), Kimi K2.5 (Team, 2026), GLM 4.5V-3.06B (Team et al., 2025), and Ministral Small 2603 (Liu et al., 2026a). We additionally report a **human baseline** evaluation (§F.3) to contextualize model performance relative to human reasoning ability.

Evaluation Process and Metrics. Models generate free-form responses, which are evaluated using an LLM-based judge under a structured prompting scheme. Full prompt

Table 2. Open-ended Evaluation Accuracy (%)(↑) across spatial reasoning Categories.

Model	Phys. (#1312)	Pos.Geo. (#2596)	Knowledge (#2332)	Re.Lo.Ca. (#455)	Temporal (#218)	Quant. (#503)	Percept. (#422)	Overall (#7838)
Proprietary Models								
Claude 4.5 Haiku	31.48	29.75	49.23	47.91	53.21	17.30	42.89	37.45
Gemini-2.5-Flash	35.90	35.68	40.93	49.45	41.74	11.55	45.02	36.99
Gemini-3.1-Flash-Lite	**46.49**	**42.22**	**52.62**	49.89	50.92	15.51	**49.76**	**45.20**
GPT-4o-mini	26.37	24.10	23.28	42.86	23.39	13.52	41.23	25.42
GPT-5-mini	38.87	35.17	52.60	**55.60**	**54.13**	**16.10**	46.92	41.95
o4-mini	37.50	35.09	47.83	55.82	55.50	16.10	45.50	40.32
Open-Source Models								
Gemma-3-4B-it	23.70	23.04	11.54	32.53	10.09	8.15	36.26	19.66
Qwen3-VL-8B-Instruct	34.53	34.78	31.40	44.62	45.41	15.11	40.28	33.45
Qwen3.5-9B	23.34	23.42	21.83	31.87	35.32	7.75	36.73	23.10
Llama-3.2-11B-Vision-Ins.	23.78	24.85	17.58	37.14	23.39	8.75	34.60	22.54
Gemma-4-31B-it	**44.59**	33.20	25.56	43.74	48.17	15.31	**49.53**	33.50
Aya-Vision-32B	24.07	21.93	16.90	35.46	26.61	15.31	40.52	22.89
Qwen3.6-35B-A3B	35.31	33.22	39.38	45.81	31.13	11.33	40.67	34.84
Kimi K2.5	33.08	31.97	34.32	39.07	36.70	10.54	42.18	32.45
GLM-4.5V	30.95	**37.30**	**34.06**	**50.77**	**44.50**	**15.31**	47.74	**35.29**
Mistral Small 2603	26.07	24.50	29.42	36.48	50.92	14.31	32.70	27.32
Human Baseline	87.35	89.10	94.17	94.51	91.28	95.23	91.71	91.22

details and evaluation protocols are provided in §F. To validate the reliability of the LLM judge, we conduct an LLM–human agreement analysis (§F), comparing model judgments against human annotations using accuracy and Cohen's κ. Table 10 shows that the judge achieves perfect agreement with the human majority vote ($\kappa = 1.0$, accuracy $= 1.0$), while human annotators exhibit substantial but imperfect agreement (average pairwise $\kappa = 0.81$). These results indicate that the LLM judge reliably reproduces the aggregated human decision boundary, supporting its use for scalable evaluation. Additional experimental details, including generation settings and evaluation procedures, are provided in §F. Prompts are available in Appendix **??**.

Hard-Case Construction and Agent-Based Evaluation Setup. We construct a hard-case evaluation set consisting of instances that remain unsolved by all evaluated frontier VLMs. For each task category, we sample 25 such cases, resulting in a balanced and adversarial subset that explicitly targets current model failure modes. This design ensures that improvements measured on this set reflect genuine progress on previously intractable problems rather than overlap with standard benchmark capabilities. We evaluate two agentic systems (§F.4): (i) **SpatioXolver** (Wasi et al., 2026), a multi-agent framework that decomposes spatial reasoning into specialized roles, and (ii) a **Claude Code-based agent** (Anthropic, 2026), which performs sequential reasoning with programmatic and tool-assisted execution. Both systems receive identical inputs under the same evaluation constraints to ensure fair comparison, with reported default setup. The primary difference lies in their reasoning strategy, with **SpatioXolver** emphasizing parallel decomposition while the Claude-based agent relies on iterative, code-driven reasoning.

5. Findings and Analysis

5.1. Overall Results

The results in Table 2 reveal a substantial performance gap between current VLMs and human-level spatial reasoning, with the human baseline achieving 91.22% overall compared to a best model score of 45.20% from GEMINI-3.1-FLASH-LITE. This gap is consistent across all categories, indicating that limitations are systemic rather than task-specific. Among proprietary models, GEMINI-3.1-FLASH-LITE and GPT-5-MINI dominate, with strong performance in knowledge-grounded and relational reasoning, suggesting that scaling and alignment improve structured reasoning more than perceptual grounding. However, even top models struggle significantly on Quantitative & Comparative reasoning, where performance remains below 17% across all systems, highlighting a persistent weakness in numerical and comparative reasoning under visual contexts.

A clear trend emerges where relational and temporal reasoning tasks (*Re.Lo.Ca.*, *Temp.Dyn.*) show relatively higher performance across models (e.g., GPT-5-MINI, O4-MINI), suggesting that LLM-derived reasoning capabilities transfer partially to these domains. In contrast, Positional & Geometric reasoning remains a bottleneck, despite being central to spatial cognition, indicating insufficient geometric grounding in current VLM architectures. Open-source models exhibit a wider variance, with larger models such

as GEMMA-4-31B-IT and QWEN3.6-35B-A3B narrowing the gap but still trailing proprietary systems, particularly in knowledge-intensive tasks. Interestingly, GLM-4.5V achieves the strongest overall performance among open models, even surpassing some proprietary counterparts, suggesting that architectural or training differences may compensate partially for scale disparities. Finally, the uniformly low performance across several categories, coupled with the large human-model gap, indicates that current VLMs lack robust cross-category generalization. The results also suggest that improvements in one dimension (e.g., safety or alignment) do not necessarily translate to gains in spatial reasoning, reinforcing the need for more targeted architectural and training interventions. Detailed breakdown of scores are available in §H.

Spatial Reasoning Bottlenecks: From Element Identification to Relational Understanding. Our evaluation harness further analyzes spatial relation identification, spatial element identification, and overall spatial reasoning in §G. Spatial relation identification exhibits substantially lower and more variable performance (≈20–80%) than element identification, with consistent strengths in Quantitative and Knowledge categories and pronounced weaknesses in Physical and Temporal reasoning. Proprietary models show complementary specialization, where CLAUDE HAIKU 4.5 leads in Temporal and Quantitative tasks, while GEMINI-3.1-FLASH-LITE performs best in Physical reasoning, and smaller models like GPT-4O-MINI lag across all categories. Open-source models demonstrate higher variance but competitive peaks, with QWEN3.5-9B, QWEN3.6-35B-A3B, and KIMI K2.5 matching or exceeding proprietary systems, particularly in Knowledge, Quantitative, and Relational reasoning. Temporal reasoning remains the least stable category, showing the widest performance spread and indicating difficulty in modeling dynamic spatial changes. Overall, results confirm that spatial element identification is largely saturated, while the primary limitation lies in converting identified elements into accurate relational reasoning.

5.2. Error Analysis

To move beyond aggregate accuracy and localize failure points in the perception–reasoning pipeline, we conduct a qualitative analysis of all incorrect responses using our error analysis pipeline (§5.2), producing a two-level taxonomy with 7 top-level categories and 38 subcategories (§I). Each error is assigned to the earliest verifiable failure stage, ensuring attribution to proximate causes rather than downstream effects.

A clear scaling trend emerges: smaller models are dominated by *perceptual* (E1) and *spatial grounding* (E2) errors, while larger models shift toward *spatial transformation* (E3) and *multi-step reasoning* (E5) failures. This indicates that improvements in perception and grounding precede gains in compositional reasoning, which remains the primary bottleneck. In contrast, *format and refusal failures* (E7) persist across all scales, with all models producing invalid or evasive outputs on certain tasks, indicating miscalibrated response behavior independent of spatial ability.

The most diagnostic categories, E1 and E3, expose distinct limitations. E1 errors, including object recognition and scene understanding, persist across scales, suggesting that current vision encoders lack fine-grained attribute and grouping supervision. E3 errors, particularly in 2D folding and 3D rotation, remain high even in larger models (e.g., GPT-5-MINI, GEMINI-2.5-FLASH), indicating that core spatial transformations are not reliably learned under standard pretraining. Notably, 3D rotation errors appear in nearly all models, showing that viewpoint invariance does not emerge from scale alone.

Three cross-cutting patterns follow. First, early-stage errors (E1, E2) propagate and inflate downstream reasoning failures (E5, E6), motivating better decoupling of perception and reasoning in benchmarks. Second, instruction-following failures (E7) are triggered by constrained-output tasks, revealing a mismatch between training objectives and evaluation requirements. Third, language–spatial mapping errors (E6) are concentrated in egocentric settings, indicating systematic bias in perspective-taking.

These findings motivate three directions: (i) 3D-aware or simulation-based supervision for transformation tasks, (ii) calibration-focused training for constrained outputs, and (iii) explicit supervision for reference-frame alignment. Full taxonomy and per-model analyses are provided in §I.

5.3. Agent-Based Hard Case Evaluation

Agent-based decomposition recovers some instances unsolved by standalone VLMs, but gains vary by reasoning type. The Claude Code Agent performs better on knowledge-grounded and quantitative tasks, indicating stronger symbolic and retrieval capabilities. In contrast, SPATIOXOLVER (Wasi et al., 2026) shows clear advantages in relational and temporal reasoning, likely due to step-wise multi-agent decomposition. Both systems perform identically on perceptual tasks, suggesting limited impact on low-level visual interpretation, and both fail on physical and geometric reasoning, exposing persistent gaps. Overall, SPATIOXOLVER achieves higher aggregate performance, while complementary strengths indicate potential for hybrid designs.

Table 3. Agent-based performance on hard-case evaluation (25 hard tasks per category)

Category	Claude Code Agent	SPATIOXOLVER	Gap (Claude Code-SPATIOXOLVER)
Knowledge-Grounded Reasoning	16.00%	12.00%	-4.00
Perceptual Understanding	20.00%	20.00%	0.00
Physical Reasoning	0.00%	0.00%	0.00
Positional & Geometric Reasoning	0.00%	0.00%	0.00
Quantitative & Comparative Reasoning	8.00%	0.00%	-8.00
Relational, Logical & Causal Reasoning	12.00%	28.00%	+16.00
Temporal & Dynamic Reasoning	4.00%	24.00%	+20.00
Overall	8.57%	12.00%	+3.43

References

Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Anthropic. Claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, October 2025.

Anthropic. Claude code, 2026. URL https://www.anthropic.com/product/claude-code. Accessed: 2026-05-05.

Azuma, D., Miyanishi, T., Kurita, S., and Kawanabe, M. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19129–19139, 2022.

Baddeley, A. Working memory. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, 321(2-3):167–173, 1998.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X.-H., Cheng, Z., Deng, L., Ding, W., Fang, R., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, Q., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L. Y., Ren, X., yi Ren, X., Song, S., chen Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Bai, X., Wang, Y., Li, D., Ping, H., and Chunxu, L. STAR: Spatial-temporal tracklet matching for multi-object tracking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=fmCnNQjZrr.

Bar-Anan, Y., Liberman, N., and Trope, Y. The association between psychological distance and construal level: evidence from an implicit association test. *Journal of experimental psychology: General*, 135(4):609, 2006.

Biswas, S., Khan, M. N. H., and Islam, B. OWL : Geometry-aware spatial reasoning for audio large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zPv46YKv3w.

Cai, H., Yang, Y., and Hu, W. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models, 2025a. URL https://arxiv.org/abs/2502.00698.

Cai, W., Ponomarenko, Y., Yuan, J., Li, X., Yang, W., Dong, H., and Zhao, B. Spatialbot: Precise spatial understanding with vision language models. *CoRR*, abs/2406.13642, 2024.

Cai, Z., Wang, Y., Sun, Q., Wang, R., Gu, C., Yin, W., Lin, Z., Yang, Z., Wei, C., Shi, X., Deng, K., Han, X., Chen, Z., Li, J., Fan, X., Deng, H., Lu, L., Li, B., Liu, Z., Wang, Q., Lin, D., and Yang, L. Has gpt-5 achieved spatial intelligence? an empirical study, 2025b. URL https://arxiv.org/abs/2508.13142.

Cao, Z., Zhu, J., Ai, H., Jiang, L., Lyu, Y., and Xiong, H. ST^{2}360d: Spatial-to-temporal consistency for training-free 360 monocular depth estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=d0bm4upnQ8.

Chabris, C. F., Jerde, T. E., Woolley, A. W., Gerbasi, M. E., Schuldt, J. P., Bennett, S. L., Hackman, J. R., and Kosslyn, S. M. Spatial and object visualization cognitive styles: Validation studies in 3800 individuals. *Group brain technical report*, 2(1-20):2, 2006.

Chang, E., Huang, Z., Liao, Y., Bhavsar, S. R., Param, A., Stark, T., Ahmadyan, A., Yang, X., Wang, J., Abdullah, A., Nguyen, G., Iyer, A., Hall, D., Li, E., Moon, S., Scheffer, N., Ahmed, K., Damavandi, B., Wanga, R., Kumar, A., Patel, R., and Dong, X. L. Wearvqa: A visual question answering benchmark for wearables in egocentric authentic real-world scenarios, 2025. URL https://arxiv.org/abs/2511.22154.

Charoenpitaks, K., Nguyen, V.-Q., Suganuma, M., Arai, K., Totsuka, S., Ino, H., and Okatani, T. Tb-bench: Training and testing multi-modal ai for understanding spatio-temporal traffic behaviors from dashcam images/videos, 2025. URL https://arxiv.org/abs/2501.05733.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L. J., and Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14455–14465. IEEE, 2024.

Chen, L., Xie, W., Liang, Y., He, H., Zhao, H., Yang, Z., Huang, Z., Wu, H., Lu, H., charles, Y., Bao, Y., Fan, Y., Li, G., Shen, H., Chen, X., Xu, W., Si, S., Cai, Z., Chai, W., Huang, Z., Liu, F., Liu, T., Chang, B., Hu, X., Chen, K., Ren, Y., Liu, Y., Gong, Y., and Li, K. Babyvision: Visual reasoning beyond language, 2026a. URL https://arxiv.org/pdf/2601.06521.

Chen, M., Cui, Z., Liu, X., Xiang, J., Zheng, C., Li, J., and Shlizerman, E. SAVVY: Spatial awareness via audio-visual LLMs through seeing and hearing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=zwCb9cKHpd.

Chen, P., Lou, Y., Cao, S., Guo, J., Fan, L., Wu, Y., Yang, L., Ma, L., and Ye, J. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026c. URL https://openreview.net/forum?id=PRCizVyL1K.

Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J. C., Geva, M., He, J., Wu, J., and Li, M. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *CoRR*, abs/2503.01773, 2025.

Chen, S., Wang, C., Xu, R., Peixingtian, yukun Song, Lin, J., Xu, W., jingyizhang, Guo, L., and Xu, S. SAGE: Spatial-visual adaptive graph exploration for efficient visual place recognition. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=DCpbEXqPvS.

Cheng, A., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. Spatialrgpt: Grounded spatial reasoning in vision-language models. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. Spatialrgpt: Grounded spatial reasoning in vision language models, 2024b. URL https://arxiv.org/abs/2406.01584.

Cheng, X., Zhang, W., Zhang, S., Yang, J., Guan, X., Wu, X., Li, X., Zhang, G., Liu, J., Mai, Y., Zeng, Y., Wen, Z., Jin, K., Wang, B., Zhou, W., Lu, Y., Li, T., Huang, W., and Li, Z. Simplevqa: Multimodal factuality evaluation for multimodal large language models, 2025. URL https://arxiv.org/abs/2502.13059.

Collaboration, O. X., Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Raj, A., Singh, A., Brohan, A., Raffin, A., Wahid, A., Burgess-Limerick, B., Kim, B., Schölkopf, B., Ichter, B., Lu, C., Xu, C., Finn, C., Xu, C., Chi, C., Huang, C., Chan, C., Pan, C., Fu, C., Devin, C., Driess, D., Pathak, D., Shah, D., Büchler, D., Kalashnikov, D., Sadigh, D., Johns, E., Ceola, F., Xia, F., Stulp, F., Zhou, G., Sukhatme, G. S., Salhotra, G., Yan, G., Schiavi, G., Kahn, G., Su, H., Fang, H., Shi, H., Amor, H. B., Christensen, H. I., Furuta, H., Walke, H., Fang, H., Mordatch, I., Radosavovic, I., and et al. Open x-embodiment: Robotic learning datasets and RT-X models. *CoRR*, abs/2310.08864, 2023.

Danish, M. S., Munir, M. A., Shah, S. R. A., Kuckreja, K., Khan, F. S., Fraccaro, P., Lacoste, A., and Khan, S. Geobench-vlm: Benchmarking vision-language models for geospatial tasks, 2024. URL https://arxiv.org/abs/2411.19325.

Dash, S., Nan, Y., Dang, J., Ahmadian, A., Singh, S., Smith, M., Venkitesh, B., Shmyhlo, V., Aryabumi, V., Beller-Morales, W., Pekmez, J., Ozuzu, J., Richemond, P. H., Locatelli, A. F., Frosst, N., Blunsom, P., Gomez, A., Zhang, I., Fadaee, M., Govindassamy, M., Roy, S., Gall'e, M., Ermiş, B. H., Ustun, A., and Hooker, S. Aya vision: Advancing the frontier of multilingual multimodality. *ArXiv*, abs/2505.08751, 2025. URL https://arxiv.org/abs/2505.08751.

Deng, N., Gu, L., Ye, S., He, Y., Chen, Z., Li, S., Wang, H., Yin, J., Wei, Q., Yang, T., Dou, M., He, T., Shao, W., Zhang, K., Wang, Y., Shi, B., Zhang, Y., Dai, J., Qiao, Y., Wang, W., and Zhang, H. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L6bEitSMeu.

Du, M., Wu, B., Li, Z., Huang, X., and Wei, Z. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pp. 346–355. Association for Computational Linguistics, 2024.

Feng, Z., Kang, Z., Wang, Q., Du, Z., Yan, J., Shubin, S., Yuan, C., Liang, H., Deng, Y., Li, Q., Yang, R., An, R., Zheng, L., Wang, W., Chen, S., Xu, S., Liang, Y., Yang, J., and Guo, B. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jXDZJAfRZB.

Fu, M., Peng, Y., Chen, D., Zhou, Z., Liu, B., Wan, Y., Zhao, Z., Yu, P. S., and Krishna, R. Seeking and updating with live visual knowledge, 2025. URL https://arxiv.org/abs/2504.05288.

Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N. A., Ma, W., and Krishna, R. BLINK: multimodal large language models can see but not perceive. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer, 2024.

Gardner, H. E. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011.

Gemma Team, Google DeepMind. Gemma 3 technical report, 2025. URL https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf. Accessed: 2025-09-11.

Gemma Team, Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-05.

Gholami, M., Rezaei, A., Weimin, Z., Mao, S., Zhou, S., Zhang, Y., and Akbari, M. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fqehqG4WvL.

Gong, Z., Li, W., Ma, X., Li, S., Wang, Z., Li, S., Ji, J., Yang, X., Luo, G., Yan, J., and Ji, R. SpaCE-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Df7UjwEgIx.

Google DeepMind. Gemini 3: The Frontier of Vision AI. https://blog.google/innovation-and-ai/technology/developers-tools/gemini-3-pro-vision/, 2025. Accessed: 2026-05-05.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A.,

Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gu, X., Shen, Y., Luo, C., Luo, T., Huang, Y., Lin, Y., Fan, H., and Zhang, L. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WOzffPgVjF.

Han, M., Yang, D., Qu, L., Chen, Z., Li, G., Wang, H., Wang, J., and Zhang, L. Fusing pixels and genes: Spatially-aware learning in computational pathology. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uVXO6gzVzj.

Hao, J., Liang, N., Luo, Z., XU, X., Zhong, W., Yi, R., Jin, Y., Lyu, Z., Zheng, F., Ma, L., and Pang, J. Mesatask: Towards task-driven tabletop scene generation via 3d spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=U88JlpY0vR.

Hosain, M. T., Rahman, S., Morol, M. K., and Parvez, M. R. Xolver: Multi-agent reasoning with holistic experience learning just like an olympiad team, 2025. URL https://arxiv.org/abs/2506.14234.

Hu, W., Hong, Y., Wang, Y., Gao, L., Wei, Z., Yao, X., Peng, N., Bitton, Y., Szpektor, I., and Chang, K.-W. 3DLLM-mem: Long-term spatial-temporal memory for embodied 3d large language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=q5QaTQcUbS.

Hu, Y., Yuan, J., Wen, C., Lu, X., and Li, X. Rsgpt: A remote sensing vision language model and benchmark, 2023. URL https://arxiv.org/abs/2307.15266.

Huang, W., Wang, C., Li, Y., Zhang, R., and Fei-Fei, L. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *CoRL*, 2024.

Huang, X., He, Q., Huang, Z., Wang, B., Li, Z., Cheng, G., Dong, Y., and Huang, X. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2510.13394.

Huang, X., He, Q., Huang, Z., Wang, B., Li, Z., Cheng, G., Dong, Y., and Huang, X. Spatial-DISE: A unified benchmark for evaluating spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bMINsPQpME.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019.

Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., and Yi, L. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models, 2025. URL https://arxiv.org/abs/2506.03135.

Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., and Yi, L. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6nZKT2rL0H.

Jiang, Y., Zhang, J., Sun, K., Sourati, Z., Ahrabian, K., Ma, K., Ilievski, F., and Pujara, J. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning, 2024. URL https://arxiv.org/abs/2404.13591.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Kamath, A., Hessel, J., and Chang, K. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 9161–9175. Association for Computational Linguistics, 2023.

Kapoor, S., Stroebl, B., Kirgis, P., Nadgir, N., Siegel, Z. S., Wei, B., Xue, T., Chen, Z., Chen, F., Utpala, S., Ndzomga, F., Oruganty, D., Luskin, S., Liu, K., Yu, B., Arora, A., Hahm, D., Trivedi, H., Sun, H., Lee, J., Jin, T., Mai, Y., Zhou, Y., Zhu, Y., Bommasani, R., Kang, D., Song, D., Henderson, P., Su, Y., Liang, P., and Narayanan, A. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vUaY1t64ZZ.

Kharlamova, A., He, B., Ma, C., and Liu, X. Spatial CAPTCHA: Generatively benchmarking spatial reasoning for human-machine differentiation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8LAc5kzjhz.

Kosoy, E., Dahmani, A., Lampinen, A. K., Comsa, I. M., Jeong, S., Dasgupta, I., and Allen, K. Decoupling the components of geometric understanding in vision language models. *CoRR*, abs/2503.03840, 2025.

Le, T., Fang, H., Li, J., Nguyen, T., Mi, L., Orsborn, A. L., Sümbül, U., and Shlizerman, E. SPINT: Spatial permutation-invariant neural transformer for consistent intracortical motor decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=XjayhGBraW.

Lee, I., Park, W., Jang, J., Noh, M., Shim, K., and Shim, B. Towards comprehensive scene understanding: Integrating first and third-person views for LVLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=m5wrqqcWbN.

Lee, J., Jung, J., Han, J., Narihira, T., Fukuda, K., Seo, J., Hong, S., Mitsufuji, Y., and Kim, S. 3d scene prompting for scene-consistent camera-controllable video generation. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=3XxoBwMusJ.

Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., ZENG, L., and Li, H. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=euMVC1DO4k.

Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., and Zhuang, Y. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=KtrFXlvgrK.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. BLIP-2: bootstrapping language-image pre-training with frozen

image encoders and large language models. *CoRR*, abs/2301.12597, 2023.

Li, J., Wang, H., chen, J., Liu, Y., Dong, Z., and Yang, B. SCot: Teaching 3d-LLMs to think spatially with million-scale cot annotations. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=5Tph6wFMOm.

Li, L., Bigverdi, M., Gu, J., Ma, Z., Yang, Y., Li, Z., Choi, Y., and Krishna, R. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=fbGmSV6tUw.

Li, P., Song, P., Li, W., Yao, H., Guo, W., Xu, Y., Liu, D., and Xiong, H. See&trek: Training-free spatial prompting for multimodal large language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026e. URL https://openreview.net/forum?id=2exr4mlbx1.

Li, T., Cao, X., Wang, Y., Yang, X., Zou, X., and Hong, B. Hippoformer: Integrating hippocampus-inspired spatial memory with transformers. In *The Fourteenth International Conference on Learning Representations*, 2026f. URL https://openreview.net/forum?id=hxwV5EubAw.

Li, W., Yu, Z., and Alahi, A. Voxdet: Rethinking 3d semantic scene completion as dense object detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026g. URL https://openreview.net/forum?id=lMhNrt0Bnm.

Li, Y., Ma, Y., Zhang, G., Yuan, R., Zhu, K., Guo, H., Liang, Y., Liu, J., Wang, Z., Yang, J., Wu, S., Qu, X., Shi, J., Zhang, X., Yang, Z., Wen, Y., Wang, Y., Li, S., Zhang, Z., Liu, Z., Benetos, E., Huang, W., and Lin, C. Omnibench: Towards the future of universal omni-language models, 2024. URL https://arxiv.org/abs/2409.15272.

Liu, A., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sad'e, A., Jeffares, A., Jiang, A. Q., Cahill, A., Gavaudan, A., Sablayrolles, A., H'eliou, A., You, A., Ehrenberg, A., Lo, A. D., Eliseev, A., Calvi, A., Sooriyarachchi, A., Bout, B., Rozière, B., Monicault, B. D., Lanfranchi, C., Barreau, C., Courtot, C., Grattarola, D., Dabert, D., de Las Casas, D., Chane-Sane, E., Ahmed, F., Berrada, G., Ecrepont, G., Guinet, G., Novikov, G. S., Kunsch, G., Lample, G., Martin, G., Gupta, G., Ludziejewski, J., Rute, J., Studnia, J., Amar, J., Delas, J., Roberts, J. S., Yadav, K., Chandu, K. R., Jain, K., Aitchison, L., Fainsin, L., Blier, L., Zhao, L., Martin, L., Saulnier, L., Gao, L., Buyl, M., Jennings, M., Pellat, M., Prins, M., Poir'ee, M., Guillaumin, M., Dinot, M., Futeral, M., Darrin, M., Augustin, M., Chiquier, M., Schimpf, M., Grinsztajn, N., Gupta, N., Raghuraman, N., Bousquet, O., Duchenne, O., Wang, P., von Platen, P., Jacob, P., Wambergue, P., Kurylowicz, P., Muddireddy, P. R., Chagniot, P., Stock, P., Agrawal, P., Torroba, Q., Sauvestre, R., Soletskyi, R., Menneer, R., Vaze, S., Barry, S., Gandhi, S., Waghjale, S., Gandhi, S., Ghosh, S., Mishra, S., Aithal, S., Antoniak, S., Scao, T. L., Cachet, T., Sorg, T. S., Lavril, T., Saada, T. N., Chabal, T., Foubert, T., Robert, T., Wang, T., Lawson, T., Bewley, T., Edwards, T., Jamil, U., Tomasini, U. M., Nemychnikova, V., Phung, V., Maladiere, V., Richard, V., Bouaziz, W., Li, W.-D., Marshall, W., Li, X., Yang, X., Ouahidi, Y. E., Wang, Y., Tang, Y., and Ramzi, Z. Ministral 3. *ArXiv*, abs/2601.08584, 2026a. URL https://arxiv.org/abs/2601.08584.

Liu, F., Emerson, G., and Collier, N. Visual spatial reasoning. *Trans. Assoc. Comput. Linguistics*, 11:635–651, 2023.

Liu, H., Zhou, L., Yang, Z., and Yao, C. Step-aware residual-guided diffusion for EEG spatial super-resolution. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=2vN2GlfV6T.

Liu, S., Gemp, I., Marris, L., Piliouras, G., Heess, N., and Lanctot, M. Re-evaluating open-ended evaluation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kbOAIXKWgx.

Liu, Y., Ma, M., Yu, X., Ding, P., Zhao, H., Sun, M., Huang, S., and Wang, D. SSR: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026c. URL https://openreview.net/forum?id=SOc0tHCewe.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Luo, Z., Zhang, C., Yong, S., Dai, C., Wang, Q., Ran, H., Shi, G., Sycara, K. P., and Xie, Y. pyspatial: Generating 3d visual programs for zero-shot spatial reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yv15C8ql24.

Ma, W., Chou, Y.-C., Liu, Q., Wang, X., de Melo, C. M., Xie, J., and Yuille, A. Spatialreasoner: Towards explicit and

generalizable 3d spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=hFaXVjRFHI.

Ma, W., Wang, C., Yuan, R., Chen, H., Dai, N., Zhou, S. K., Yang, Y., Yuille, A., and Chen, J. Causalspatial: A benchmark for object-centric causal spatial reasoning, 2026b. URL https://arxiv.org/abs/2601.13304.

Ma, X., SUN, T., Chen, S., Bhalgat, Y. S., Gu, J., Chang, A. X., Armeni, I., Laina, I., Peng, S., and Prisacariu, V. A. Do 3d large language models really understand 3d spatial relationships? In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=3vlMiJwo8b.

Ma, Y., Wu, H., Zhou, H., Weng, H., Wang, J., and Long, M. Physense: Sensor placement optimization for accurate physics sensing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026d. URL https://openreview.net/forum?id=zIzZxDsNNP.

Mao, Y., Zhong, J., Fang, C., Zheng, J., Tang, R., Zhu, H., Tan, P., and Zhou, Z. SpatialLM: Training large language models for structured indoor modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=u1lNQH5upa.

Maruf, M., Daw, A., Mehrab, K. S., Manogaran, H. B., Neog, A., Sawhney, M., Khurana, M., Balhoff, J. P., Bakis, Y., Altintas, B., Thompson, M. J., Campolongo, E. G., Uyeda, J. C., Lapp, H., Bart, H. L., Mabee, P. M., Su, Y., Chao, W.-L., Stewart, C., Berger-Wolf, T., Dahdul, W., and Karpatne, A. Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images, 2024. URL https://arxiv.org/abs/2408.16176.

OpenAI. Gpt-4v(ision) system card, 2023. URL https://openai.com/research/gpt-4v-system-card.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. Openai o3 and o4-mini system card, April 2025. URL https://openai.com/index/o3-o4-mini-system-card/. System Cards. Accessed: 2025-09-11.

OpenAI. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Palmer, S. *Vision Science: From Photons to Phenomenology*, volume 1. MIT Press, 01 1999.

Pan, Z. and Liu, H. Metaspatial: Reinforcing 3d spatial reasoning in VLMs for the metaverse. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EdQzLC0Zra.

Pothiraj, A., Stengel-Eskin, E., Cho, J., and Bansal, M. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. *CoRR*, abs/2504.15485, 2025.

Previc, F. H. The neuropsychology of 3-d space. *Psychological bulletin*, 124(2):123, 1998.

Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., and Ma, K. Shapellm: Universal 3d object understanding for embodied interaction. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIII*, volume 15101 of *Lecture Notes in Computer Science*, pp. 214–238. Springer, 2024.

Qi, Z., Zhang, W., Ding, Y., Dong, R., Yu, X., Li, J., Xu, L., Li, B., He, X., Fan, G., Zhang, J., He, J., Gu, J., Jin, X., Ma, K., Zhang, Z., Wang, H., and Yi, L. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *CoRR*, abs/2502.13143, 2025.

Qi, Z., Zhang, W., Ding, Y., Dong, R., Yu, X., Li, J., Xu, L., Li, B., He, X., Fan, G., Zhang, J., He, J., Gu, J., Jin, X., Ma, K., Zhang, Z., Wang, H., and Yi, L. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=kmv7yg6QXv.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Ramalho, T., Kociský, T., Besse, F., Eslami, S. M. A., Melis, G., Viola, F., Blunsom, P., and Hermann, K. M. Encoding spatial relations from natural language. *CoRR*, abs/1807.01670, 2018.

Roberts, J., Taesiri, M. R., Sharma, A., Gupta, A., Roberts, S., Croitoru, I., Bogolin, S.-V., Tang, J., Langer, F., Raina, V., Raina, V., Xiong, H., Udandarao, V., Lu, J., Chen, S., Purkis, S., Yan, T., Lin, W., Shin, G., Yang, Q., Nguyen, A. T., Atkinson, D. I., Baranwal, A., Coca, A., Dang, M., Dziadzio, S., Kunz, J. D., Liang, K., Lo, A., Pulfer, B., Walton, S., Yang, C., Han, K., and Albanie, S. Zerobench: An impossible visual benchmark

for contemporary large multimodal models, 2025. URL https://arxiv.org/abs/2502.09696.

Santos, J. E., Marcato, A., Colman, R., Lubbers, N., and Lin, Y. T. Discrete spatial diffusion: Intensity-preserving diffusion modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=kz3w2A2y0e.

Saura, M. N., Allmendinger, R., Pan, W., and Papamarkou, T. Spatial-aware decision-making with ring attractors in reinforcement learning systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=FthMPOhgfp.

Shen, Y., Bi, Q., Wang, Z., Yang, Z., Wang, C., Zhang, Z., Tiwari, P., Pimentel, A. D., and Pathania, A. Efficient multimodal spatial reasoning via dynamic and asymmetric routing. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=BQASoLmREU.

Shen, Y., Liu, Y., Zhu, J., Cao, X., Zhang, X., He, Y., Ye, W., Rehg, J. M., and Lourentzou, I. Fine-grained preference optimization improves spatial reasoning in VLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=L9vV3wVC72.

Shiri, F., Guo, X., Far, M., Yu, X., Haf, R., and Li, Y. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Al-Onaizan, Y., Bansal, M., and Chen, Y. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 21440–21455. Association for Computational Linguistics, 2024.

Song, C. H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., and Birchfield, S. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *CoRR*, abs/2411.16537, 2024.

Sterz, H., Pfeiffer, J., and Vulić, I. Dare: Diverse visual question answering with robustness evaluation, 2024. URL https://arxiv.org/abs/2409.18023.

Sturm, D., Bensalem, H., and Sbalzarini, I. F. Spatially informed autoencoders for interpretable visual representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=09YSBymX6O.

Szymanska, E., Dusmanu, M., Buurlage, J., Rad, M., and Pollefeys, M. Space3d-bench: Spatial 3d question answering benchmark. *CoRR*, abs/2408.16662, 2024.

Tang, K., Gao, J., Zeng, Y., Duan, H., Sun, Y., Xing, Z., Liu, W., Lyu, K., and Chen, K. Lego-puzzles: How good are mllms at multi-step spatial reasoning?, 2025. URL https://arxiv.org/abs/2503.19990.

Team, G. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Team, K. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Xu, J., Zhu, J., Chen, J., Chen, J., Chen, J., Lin, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Yang, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Tu, S., Meng, S., Zhang, T., Luo, T., Hao, T., Tong, T., Li, W., Jia, W., Liu, X., Zhang, X., Lyu, X., Fan, X., Huang, X., Wang, Y., Xue, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Shi, Y., Huang, Y., Niu, Y., Wang, Y., Yue, Y., Li, Y., Zhang, Y., Wang, Y., Wang, Y., Zhang, Y., Xue, Z., Hou, Z., Du, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., and Tang, J. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Trope, Y. and Liberman, N. Construal-level theory of psychological distance. *Psychological review*, 117(2):440, 2010.

Wang, J., Zheng, Z., Chen, Z., Ma, A., and Zhong, Y. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering, 2023. URL https://arxiv.org/abs/2312.12222.

Wang, J., Yang, X., Sun, K., Suresh, P., Sharma, S., Czyzewski, A., Andersen, D., Appini, S., Banerjee, A., Choudhary, S., Ghasemlou, S., Guan, Z., Iyer, A., Khan, H., Kong, L., Luo, R., Ma, T., Qiao, Z., Tran, D., Xu, W., Yeatman, S., Zhou, C., Gujral, G., Xia, Y., Moon, S., Scheffer, N., Shah, N., Chang, E., Liu, Y., Metze, F., Stark, T., Feizollahi, Z., Jessee, A., Pujari, M., Aly, A., Damavandi, B., Wanga, R., Kumar, A., Patel, R., tau Yih, W., and Dong, X. L. Crag-mm: Multi-modal multi-turn comprehensive rag benchmark, 2025a. URL https://arxiv.org/abs/2510.26160.

Wang, K., Pan, J., Shi, W., Lu, Z., Zhan, M., and Li, H. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. URL https://arxiv.org/abs/2402.14804.

Wang, Q., Yin, B., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., Xie, S., Wu, J., Fei-Fei, L., and Li, M. Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=0FhrtdKLtD.

Wang, S., Liang, C., Gao, Y., Yu, E., Li, S., Li, J., and Wang, H. Cityseeker: How do VLMs explore embodied urban navigation with implicit human needs? In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=hzf23XSDcs.

Wang, S., Pei, M., Sun, L., Deng, C., Li, Y., Shao, K., Tian, Z., Zhang, H., and Wang, J. Spatialviz-bench: A cognitively-grounded benchmark for diagnosing spatial visualization in MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=OqZ7bm28Xx.

Wang, X., Ma, W., Zhang, T., de Melo, C. M., Chen, J., and Yuille, A. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models, 2025b. URL https://arxiv.org/abs/2502.08636.

Wang, Z., Qin, Y., Zeng, L., and Zhang, R. High-dynamic radar sequence prediction for weather nowcasting using spatiotemporal coherent gaussian representation. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=Cjz9Xhm7sI.

Wang, Z., Lee, S., and Lee, G. H. Dynam3d: Dynamic layered 3d tokens empower VLM for vision-and-language navigation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026d. URL https://openreview.net/forum?id=s6k9l5yX8e.

Wasi, A. T., Faisal, W., Rahman, A., Anik, M. A., Shahriar, M., Topu, M. M., Meem, S. T., Priti, R. N., Mitu, S. A., Hoque, M. I., Ridoy, S. Z., Ali, M. E., Hawasly, M., Raza, M., and Parvez, M. R. Spatialab: Can vision–language models perform spatial reasoning in the wild? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fWWUPOb0CT.

Wei, T., Lin, Y., Zhou, Z., Wen, H., Hu, J., Guo, S., Lin, Y., Cong, G., and Wan, H. Transfertraj: A vehicle trajectory learning model for region and task transferability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=XQ87Vo9GIz.

Wu, D., Liu, F., Hung, Y.-H., and Duan, Y. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=RnXS7aK4rK.

Wu, G., Su, B., Zhao, Y., Wang, P., Lin, Y., and Yang, H. F. Towards physics-informed spatial intelligence with human priors: An autonomous driving pilot study. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=pEUBqS8nTk.

Xu, R., Lu, D., Zhao, Z., Tan, X., Wang, X., Yuan, S., Chen, J., and Yinghui, X. ORIGAMISPACE: Benchmarking multimodal LLMs in multi-step spatial reasoning with mathematical constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=y7ahj9RoXQ.

Xu, S., Dou, Z., Shi, M., Pan, L., Ho, L., Wang, J., Liu, Y., Lin, C., Ma, Y., Wang, W., and Komura, T. Mospa: Human motion generation driven by spatial audio. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=X2r9D46kvI.

Xu, Y., Zhu, L., and Yang, Y. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms, 2024. URL https://arxiv.org/abs/2410.12332.

Yan, Y., Zhou, Z., Gao, X., Li, G., Li, S., Chen, J., Pu, Q., and Pu, J. Learning spatial-aware manipulation ordering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=iM8154lJCx.

Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L., and Xie, S. Thinking in space: How multimodal large language models see, remember, and recall spaces. *CoRR*, abs/2412.14171, 2024.

Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., Lin, D., Wang, T., and Pang, J. Mmsi-bench: A benchmark for multi-image spatial intelligence, 2025. URL https://arxiv.org/pdf/2505.23764.

Yang, S., Yang, J., Huang, P., II, E. L. B., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Fergus, R., LeCun, Y., Fei-Fei, L., and Xie, S. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rBFDvZu6pb.

Ye, J., Wang, F., Gao, N., Yu, J., Yangkun, Z., Wang, B., Zhang, J., Jin, W., Fu, Y., Zheng, F., Chen, Y., and Pang, J. Spatially guided training for vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eKhOrQWAVJ.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. URL https://arxiv.org/abs/2308.02490.

Yuan, Y., Cui, H., Chen, Y., Dong, Z., Ni, F., Kou, L., Liu, J., Li, P., ZHENG, Y., and HAO, J. From seeing to doing: Bridging reasoning and decision for robotic manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yngvAamNQi.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2023. URL https://arxiv.org/abs/2311.16502.

Zeng, S., Qi, D., Chang, X., Xiong, F., Shichao, X., Wu, X., Liang, S., Xu, M., and Wei, X. JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RnuB0Nlbd5.

Zhan, X., Huang, W., Sun, H., Fu, X., Ma, C., Cao, S., Jia, B., Lin, S., Yin, Z., BAI, L., Ouyang, W., Li, Y., Guo, J., and Guo, Y. Actial: Activate spatial reasoning ability of multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=jquTBzt3Av.

Zhang, H., Liu, M., Li, Z., Wen, H., Guan, W., Wang, Y., and Nie, L. Spatial understanding from videos: Structured prompts meet simulation data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=SBYCu5uJJf.

Zhang, S., Lian, Z., Yang, J., Li, D., Pang, G., Liu, F., Han, B., Li, S., and Tan, M. Physics-driven spatiotemporal modeling for AI-generated video detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=HiBoJLCyEo.

Zhang, Y., Corcodel, R., Hori, C., Cherian, A., and Zhao, D. Spinbench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=r7rUDgGYC4.

Zhang, Z., Hu, F., Lee, J., Shi, F., Kordjamshidi, P., Chai, J., and Ma, Z. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=84pDoCD4lH.

Zhang, Z., Li, H., Dai, Y., Zhu, Z., Zhou, L., Liu, C., Wang, D., Tay, F. E. H., Chen, S., Liu, Z., Liu, Y., Li, X., and Zhou, P. From spatial to actions: Grounding vision-language-action model in spatial foundation priors. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=fzmittHfq3.

Zhu, F., Wang, H., Xie, Y., Gu, J., Ding, T., Yang, J., and Jiang, H. Struct2d: A perception-guided framework for spatial reasoning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=rbIlWjTFKj.

Zhu, Z., LIANG, K., Meng, L., Liu, M., Liu, S., Guan, R., Li, M., Liu, W., and Liu, X. SAINT: Sequence-aware integration for spatial transcriptomics multi-view clustering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=EBHZNmygTT.

Ethics Statement

We have taken multiple steps to ensure ethical compliance throughout the development of this work. All images used in this benchmark are sourced from previously published or publicly available datasets, ensuring adherence to established usage and licensing norms. Where applicable, associated metadata was removed to protect privacy, and personally identifiable features (e.g., human faces or license plates) were minimized or excluded. The dataset was carefully curated with attention to copyright and licensing considerations, though minor oversights may occur despite best efforts. The benchmark construction process emphasizes transparency and fairness, with care taken to avoid disproportionate or harmful representation of any demographic group. We adhere to the NeurIPS Code of Ethics and confirm that no human subjects or sensitive personal data were collected or introduced as part of this work.

A. Related Work

Spatial reasoning in vision–language models (VLMs) sits at the intersection of perception, cognition, and grounded action. This section surveys the breadth of prior work, organized into seven thematic threads: (i) the foundational challenge of spatial reasoning in VLMs; (ii) benchmark design and its evolution; (iii) the persistent tension between scale and taxonomic coverage; (iv) evaluation format and diagnostic depth; (v) recent advances in 3D and embodied spatial understanding; (vi) emerging directions in spatial audio, transcriptomics, and cross-modal sensing; and (vii) the positioning of FRONTIERSPATIAL within this landscape. Table 4 provides a structured comparison of representative benchmarks.

A.1. Spatial Reasoning in Vision–Language Models

Spatial reasoning—the capacity to interpret geometric structures, relational layouts, and perspective-consistent representations in visual scenes—is central to embodied intelligence and a key bottleneck for real-world deployment of artificial agents (Huang et al., 2024; Qi et al., 2024; 2025). Early multimodal benchmarks such as CLEVR (Johnson et al., 2016) and GQA (Hudson & Manning, 2019) probed compositional reasoning through synthetic or template-generated scenes, enabling controlled analysis but offering limited ecological validity. With the rise of large VLMs (Li et al., 2023; Alayrac et al., 2022; OpenAI, 2023), spatial reasoning has been revisited in more naturalistic settings, yet systematic weaknesses persist. Despite strong performance on captioning and visual QA, these models routinely fail at depth ordering, occlusion inference, and consistent reference-frame alignment (Chen et al., 2025; Kosoy et al., 2025; Pothiraj et al., 2025).

Crucially, scaling alone does not resolve these deficits—structural spatial understanding requires capabilities that go beyond what language-grounded pattern matching provides (Cai et al., 2025b). Recent work has identified further failure modes: (Zhang et al., 2025) show through the COMFORT benchmark that current VLMs handle spatial frames of reference poorly, especially across languages and in ambiguous contexts, while (Zhang et al., 2026c) reveal strong egocentric bias and poor rotational understanding across 43 VLMs via the SpinBench evaluation. A recurring finding is that even state-of-the-art proprietary models leave large gaps relative to human performance: (Wang et al., 2026a) report near-random performance on spatial mental modeling tasks in the MindCube benchmark, and (Huang et al., 2026) characterize consistent gaps across 32 VLMs using the Spatial-DISE taxonomy, which partitions spatial reasoning into intrinsic/extrinsic and static/dynamic quadrants.

Efforts to close these gaps have taken multiple directions. (Li et al., 2026b) train a 3B model progressively from perception to reasoning, surpassing GPT-4o by 20.8% on spatial benchmarks via SpatialLadder. (Ma et al., 2026a) demonstrate that explicitly sharing 3D representations across perception, computation, and reasoning stages outperforms Gemini 2.0 by 9.2% on 3DSRBench. From a cognitive neuroscience perspective, (Li et al., 2026f) integrate a hippocampus-inspired spatial memory module with Transformers, achieving superior 2D/3D spatial prediction. Meanwhile, (Kharlamova et al., 2026) introduce Spatial CAPTCHA as a diagnostic tool, showing that leading MLLMs solve these geometric verification challenges at only 31% accuracy compared to near-perfect human performance—underscoring the magnitude of the remaining gap.

A.2. Spatial Reasoning Benchmarks: Coverage and Design

A growing number of benchmarks have been proposed to evaluate spatial reasoning in VLMs, yet each addresses only a narrow slice of the problem space (see Table 4). Among early embodied QA datasets, SCANQA (Azuma et al., 2022) grounded questions in 3D indoor reconstructions but was limited in scale and task variety. VISUAL SPATIAL (Liu et al., 2023) and WHAT'S UP (Kamath et al., 2023) extended coverage to household and real-world contexts, yet relied on template-driven question generation that narrows linguistic variability and overrepresents simple binary spatial relations.

EMBSPATIAL-BENCH (Du et al., 2024) and SPACE3D-BENCH (Szymanska et al., 2024) increased task variety but remained constrained in scale (2.2K and 211 samples, respectively) and annotation style.

More recent benchmarks target specific sub-skills. MMSI-BENCH (Yang et al., 2025), SPATIAL-DISE (Huang et al., 2025), Q-SPATIAL-BENCH (Cheng et al., 2024b), and CAUSALSPATIAL (Ma et al., 2026b) focus on metric depth, positional relations, or causal inference, respectively, but none provides unified coverage across the full spectrum of spatial cognition. SPATIALAB (Wasi et al., 2026;?) advances the field with 30 subcategories, manual puzzle-style annotations, and open-ended evaluation across 1,400 questions with a human accuracy of 87.57% compared to 54.93% for the best models, yet its scope leaves the holistic picture incomplete.

Broader coverage benchmarks have appeared in the recent literature. (Jia et al., 2026) introduce OmniSpatial, a cognitively-grounded benchmark covering dynamic, logical, interactive, and perspective-taking spatial reasoning across 50 subcategories. (Gong et al., 2026) propose SpaCE-10, which benchmarks compositional spatial intelligence across 10 atomic and 8 compound capabilities using real indoor scenes with point cloud input. (Wang et al., 2026c) evaluate spatial visualization across 27 MLLMs and 12 tasks with SpatialViz-Bench, notably finding that chain-of-thought prompting paradoxically degrades accuracy in open-source models. (Li et al., 2026d) present STARE, a benchmark evaluating multi-step visual simulation tasks such as 3D folding and tangram manipulation, where models approach random chance. Compositional and perspective-taking reasoning is further assessed by SpinBench (Zhang et al., 2026c), which identifies strong egocentric bias, and by the Spatial CAPTCHA framework (Kharlamova et al., 2026).

In temporally-grounded settings, (Gu et al., 2025) introduce TA-STVG, which adaptively generates object queries from video-text pairs using temporal and spatial cues for better spatio-temporal video grounding. Multi-view spatial reasoning is targeted by (Feng et al., 2026) through MV-RoboBench, which evaluates robotic manipulation settings and finds state-of-the-art models far below human performance, while (Gholami et al., 2026) introduce Ego3D-Bench for ego-centric multi-view outdoor spatial reasoning paired with the Ego3D-VLM model that improves performance via cognitive map generation.

A.3. Scale versus Taxonomic Coverage

A persistent tension in spatial benchmark design is between scale and taxonomic coverage. Large-scale datasets such as SPATIALVLM (Chen et al., 2024) provide billions of QA pairs but cover only two coarse spatial categories, rendering them unsuitable for fine-grained diagnostic evaluation. ROBOSPATIAL (Song et al., 2024) scales to one million robot-centric scenes but relies on template-based annotation that enables heuristic shortcuts and introduces a significant sim-to-real gap. High-quality manual datasets such as VSI-BENCH (Yang et al., 2024) achieve annotation fidelity but remain too small (288 scenes) for robust cross-model comparison.

Efforts to build large-scale, diverse spatial instruction corpora have gained traction. (Deng et al., 2026) introduce InternSpatial, the largest open-source spatial reasoning dataset to date, comprising 12M QA pairs across 19 instruction formats for both single- and multi-view settings. (Li et al., 2026c) contribute SCoT, a million-scale chain-of-thought annotation suite for 3D-LLMs covering perception, analysis, and planning levels. (Ma et al., 2026c) demonstrate through Real-3DQA that existing 3D-LLMs rely heavily on textual shortcuts, and propose a 3D-reweighted training objective that substantially improves genuine spatial reasoning. Despite these advances, no existing benchmark simultaneously achieves broad taxonomic coverage, sufficient scale, visual diversity, and multi-format evaluation—a gap that FRONTIERSPATIAL directly addresses.

Benchmarks targeting specific domains contribute valuable task diversity but are too narrow to inform general spatial reasoning capabilities. Remote sensing is probed by RSIEVAL (Hu et al., 2023) and GEOBENCH-VLM (Danish et al., 2024); biological imaging by VLM4BIO (Maruf et al., 2024); and wearable contexts by WEARVQA (Chang et al., 2025). Even broad-coverage benchmarks such as MMMU (Yue et al., 2023) and OMNIBENCH (Li et al., 2024) treat spatial tasks as incidental rather than primary, yielding only aggregate signals. (Yang et al., 2026) introduce Cambrian-S with the VSI-SUPER benchmark and a predictive sensing model that outperforms proprietary baselines, while (Wang et al., 2026b) benchmark VLMs on implicit-need urban navigation through CitySeeker, with top models achieving only 21.1% task completion across 8 cities.

A.4. Evaluation Formats and Diagnostic Depth

Beyond task coverage, evaluation format shapes what weaknesses are measurable. Most prior benchmarks employ closed-ended multiple-choice or binary judgments (Du et al., 2024; Kamath et al., 2023; Cai et al., 2025a; Fu et al., 2024), which

simplify scoring but allow models to exploit option elimination rather than produce genuine spatial reasoning. Benchmarks such as MM-VET (Yu et al., 2023), MARVEL (Jiang et al., 2024), and SPATIALAB (Wasi et al., 2026) introduce open-ended evaluation, but few couple this with systematic error analysis across a detailed task hierarchy.

Structured diagnostic studies—examining failure modes by task type, image complexity, and reasoning depth—are largely absent from the literature. The Spatial Intelligence Grid (SIG) schema introduced by (Wu et al., 2026b) represents a step in this direction: it encodes scene layouts and physical priors as a structured complement to text, enabling more faithful and measurable visual-spatial intelligence evaluation. Similarly, (Lee et al., 2026a) combine egocentric and exocentric views via the E3VQA benchmark and M3CoT prompting method to give LVLMs richer scene context for more accurate multi-view spatial reasoning. (Luo et al., 2026) introduce pySpatial, a zero-shot visual programming framework that equips MLLMs with 3D spatial tools via Python code generation, allowing direct inspection of model reasoning chains.

Reasoning efficiency has also emerged as an evaluation concern. (Shen et al., 2026a) show that DARE can dynamically prune multimodal tokens with asymmetric modality compression, cutting FLOPs by 40% while preserving spatial reasoning performance—demonstrating that spatial capability need not come at prohibitive computational cost. The question of *how* models reason spatially, rather than merely whether they succeed, motivates a growing body of work on chain-of-thought and structured prompting: (Zhang et al., 2026a) combine SpatialMind structured prompting with ScanForgeQA simulation-based fine-tuning to enhance 3D spatial reasoning without architectural changes, while (Zhu et al., 2026a) use bird's-eye-view images with object-centric metadata as structured 2D prompts to enable strong 3D reasoning without explicit 3D input.

A.5. 3D Spatial Understanding and Scene Reasoning

Grounding spatial reasoning in explicit 3D representations has been a central thrust of recent work. (Wu et al., 2026a) propose Spatial-MLLM, a dual-encoder architecture pairing a semantic encoder with a geometry-focused 3D spatial encoder to boost spatial reasoning from purely 2D video inputs. (Mao et al., 2026) fine-tune a standard multimodal LLM on 12K synthetic indoor scenes to generate structured 3D outputs including walls, doors, windows, and oriented bounding boxes. (Liu et al., 2026c) convert raw depth data into structured textual rationales distilled into compact latent embeddings for plug-and-play spatial reasoning enhancement. (Chen et al., 2026c) introduce depth positional encoding and a 700K-pair spatial measurement dataset (MSMU) to give VLMs strong quantitative 3D spatial measuring capabilities.

Scene-level 3D reasoning for navigation and manipulation has attracted substantial attention. (Qi et al., 2026) introduce semantic orientation defined via natural language, building the OrienText300K dataset to enable 6-DoF spatial reasoning and robotic manipulation without predefined reference frames. (Hao et al., 2026) generate task-aligned tabletop scenes via a spatial reasoning chain decomposing instructions into object inference, relation reasoning, and 3D scene graph construction. (Wang et al., 2026d) build dynamic layered 3D patch-instance-zone representations from RGB-D inputs for vision-and-language navigation, achieving state-of-the-art results on R2R-CE, REVERIE-CE, and NavRAG-CE. (Hu et al., 2026) introduce 3DLLM-Mem, which uses working memory tokens to selectively query episodic spatial-temporal memory for long-horizon embodied reasoning, outperforming the strongest baselines by 16.5% on 3DMem-Bench.

Semantic scene completion benefits from spatial reformulations as well. (Li et al., 2026g) reformulate semantic scene completion as dense object detection by decoupling voxel-level prediction into offset regression and semantic classification, achieving state-of-the-art results on SemanticKITTI. For 3D question answering, (Ma et al., 2026c) show that genuine spatial reasoning is being sidestepped by textual cues in existing 3D-LLMs, and (Shen et al., 2026b) propose SpatialReasoner-R1, combining multi-model Monte Carlo Tree Search for chain-of-thought data generation with fine-grained DPO for segment-level spatial reward—setting a new state-of-the-art on SpatialRGPT-Bench. Spatial prompting for purely vision-based understanding is addressed by (Li et al., 2026e) through See&Trek, a training-free framework combining keyframe sampling with visual trajectory simulation to improve spatial understanding from vision alone.

A.6. Embodied and Robotic Spatial Reasoning

Robotic manipulation and navigation demand spatial reasoning that operates under real-world constraints of partial observability, dynamic scenes, and grounded action. Several recent lines of work bridge VLMs with robotic spatial grounding. (Ye et al., 2026) introduce SP-VLA, which bridges VLMs with robotic control via spatial grounding pretraining and spatially-guided action post-training, achieving state-of-the-art results on SimplerEnv. (Yuan et al., 2026) propose FSD, which generates spatial relationship intermediate representations to guide robotic manipulation, achieving a 40.6% success rate on SimplerEnv. (Zhang et al., 2026d) introduce FALCON, which injects 3D spatial tokens from foundation models into a dedicated action head, achieving state-of-the-art VLA performance across both simulation and real-world tasks. (Li et al.,

2026a) propose Spatial Forcing, which aligns VLA intermediate embeddings with pretrained 3D model representations to improve spatial awareness implicitly, without requiring explicit 3D inputs at inference time.

For object ordering in cluttered scenes, (Yan et al., 2026) introduce OrderMind, which learns spatial-aware manipulation ordering via a spatial graph encoder and VLM-guided priority labeling, evaluated on a 163K-sample benchmark. Vision-and-language navigation is addressed by (Zeng et al., 2026), who decouple spatial-geometric and visual-semantic memory into dual implicit neural representations in JanusVLN for more efficient navigation. Scene layout generation benefits from (Pan & Liu, 2026)'s MetaSpatial, which uses reinforcement learning with physics-aware rewards for real-time 3D scene layout generation without post-processing. For video generation conditioned on camera control, (Lee et al., 2026b) introduce 3DScenePrompt with dual spatio-temporal conditioning and dynamic SLAM-based scene memory for scene-consistent, camera-controllable video generation.

Spatial reasoning is also critical for place recognition. (Chen et al., 2026d) introduce SAGE, which enhances visual place recognition via adaptive geo-visual graph construction and greedy clique expansion sampling with a frozen DINOv2 backbone. Cross-view spatial reasoning is activated in MLLMs by (Zhan et al., 2026) through Actial's Viewpoint Learning, using a 100K cross-view dataset with two-stage SFT and GRPO fine-tuning. Multi-view benchmarking is addressed by (Feng et al., 2026) through MV-RoboBench, while audio-visual spatial tracking is handled by (Chen et al., 2026b) through SAVVY, which tracks object trajectories and constructs a global dynamic map for 3D spatial QA across dynamic scenes with spatial audio.

A.7. Temporal, Spatial–Temporal, and Weather Prediction

Spatial reasoning extends beyond static scenes to temporally dynamic contexts. (Wang et al., 2025c) introduce the STC-GS + GauMamba framework for efficient 3D radar sequence prediction for weather nowcasting via spatiotemporal Gaussian representations. For vehicle trajectory learning, (Wei et al., 2026) propose TransferTraj, which unifies multi-modal trajectory representations with a task-transferable masking scheme for zero-shot and few-shot vehicle trajectory learning across regions and tasks.

Multi-object tracking under occlusion is addressed by (Bai et al., 2026), who propose STAR's spatial-temporal tracklet clip graph matching, linking confidently associated trajectory segments via appearance and topology features. Depth estimation for omnidirectional imagery is handled by (Cao et al., 2026) through $ST^2$360D, a training-free pipeline converting 360-degree images into perspective video sequences and exploiting temporal consistency from video depth models for coherent 360-degree depth maps. Detection of AI-generated videos is approached spatiotemporally by (Zhang et al., 2026b) through NSG-VD, which measures deviations from physical probability flow conservation using a normalized spatiotemporal gradient statistic derived from diffusion models.

A.8. Spatial Audio, Brain–Computer Interfaces, and Neuroscience

Spatial reasoning is increasingly studied in modalities beyond vision. (Biswas et al., 2026) introduce OWL, which aligns binaural audio with 3D geometry for audio LLMs, enabling clock-level azimuth estimation with geometry-aware chain-of-thought reasoning. (Xu et al., 2026b) introduce MOSPA, the first spatial audio-driven human motion dataset, paired with a diffusion-based framework for generating realistic body motion conditioned on spatial audio signals.

In brain–computer interfaces, (Le et al., 2026) introduce SPINT, a permutation-invariant Transformer for intracortical BCI decoding that uses context-dependent positional embeddings to generalize across sessions without test-time retraining. For EEG spatial super-resolution, (Liu et al., 2026b) propose SRGDiff, a step-aware residual-guided diffusion model achieving up to 40% gains over strong baselines. Deep reinforcement learning agents benefit from spatial representations as well: (Saura et al., 2026) integrate ring attractor neural structures into RL agents to encode spatial action representations and stabilize action selection, achieving a 53% improvement on Atari 100k.

A.9. Spatial Transcriptomics and Biological Spatial Modeling

Spatial reasoning also plays a role in biological and medical imaging contexts. (Zhu et al., 2026b) introduce SAINT, which augments spatial transcriptomics clustering with DNA sequence embeddings from a Nucleotide Transformer to improve spot-level clustering across tissue benchmarks. (Han et al., 2026) propose STAMP, which aligns pathology images with spatially-resolved gene expression via contrastive learning, trained on the largest Visium dataset (SpaVis-6M). For spatial organization in microscopy images, (Sturm et al., 2026) introduce SI-VAE, which incorporates point-process likelihoods

into variational autoencoders for interpretable spatial organization learning. Physics sensing in spatially distributed systems is addressed by (Ma et al., 2026d) through PhySense, which jointly learns physical field reconstruction and sensor placement optimization via a generative model with projected gradient descent.

Discrete generative modeling with spatial structure is advanced by (Santos et al., 2026) through Discrete Spatial Diffusion (DSD), a discrete-state jump process framework that strictly preserves particle counts under mass conservation constraints.

A.10. Mathematical Spatial Reasoning and Cognitive Benchmarks

Spatial reasoning that requires explicit mathematical constraint handling is an emerging frontier. (Xu et al., 2026a) introduce ORIGAMISPACE, which benchmarks MLLMs on multi-step spatial reasoning and mathematical constraint handling through origami crease pattern tasks, including RL-based code generation evaluation. The challenge of perspective-taking and mental rotation is specifically examined by (Wang et al., 2026a) through MindCube, where a "map-then-reason" approach combining spatial map generation with reinforcement learning boosts accuracy from 37.8% to 70.7%. (Li et al., 2026d) similarly probe multi-step visual simulation via STARE, finding that models approach random chance on tasks requiring mental simulation.

A.11. Positioning of FRONTIERSPATIAL

As illustrated in Table 4 and summarized numerically in Table **??**, existing benchmarks each cover at most a handful of spatial dimensions, leaving large blind spots in the evaluation landscape. The fragmentation spans multiple axes: narrow task scope (most benchmarks address fewer than five spatial sub-categories), limited scale (many high-quality datasets contain fewer than 2K samples), constrained visual diversity (synthetic or template-heavy sources), homogeneous evaluation format (closed-ended only), and absence of automated diagnostic tooling.

FRONTIERSPATIAL addresses this fragmentation directly. By consolidating and extending 23 existing datasets through an automated curation pipeline, FRONTIERSPATIAL assembles 7,838 questions paired with 11,825 images under a four-level task hierarchy spanning 7 categories, 19 sub-categories, and 65 fine-grained tasks, achieving full coverage across all 22 spatial dimensions identified in our taxonomy. Unlike template-heavy predecessors, the benchmark draws from diverse real-world visual sources and spans both multiple-choice and open-ended evaluation formats.

Complementing the benchmark, we contribute an automated evaluation harness for reproducible model comparison and an automated error analysis harness for structured, fine-grained diagnostics—capabilities that prior spatial benchmarks have not provided at this scope. Together, these contributions establish FRONTIERSPATIAL as the first holistic framework for spatial reasoning evaluation in the multimodal era, addressing gaps that no single prior benchmark—whether broad (OmniSpatial (Jia et al., 2026), Spatial-DISE (Huang et al., 2026)) or deep (VSI-Bench (Yang et al., 2024), SpatiaLab (Wasi et al., 2026))—has been able to close.

Table 4. Comparison of Benchmarks Across Reasoning and Perception Dimensions

Benchmark	Predictive Physics	Temporal Reasoning	Textual Perception	3D Reasoning	Geometric Reasoning	Mathematical Reasoning	Causal Reasoning	Counting	Domain Specific	Creative & Abstract	Navigation & Control	Object-Level Perception	2D Reasoning	Relations	Motion Reasoning	Positional Relations	Measurement	Localization	Scene-Level Perception	Physical Interaction	Logical Reasoning	Comparison
OmniBench (Li et al., 2024)			✓		✓	✓			✓	✓									✓			
Spatial457 (Wang et al., 2025b)				✓	✓									✓		✓		✓				✓
MathVision (Wang et al., 2024)					✓	✓			✓				✓				✓					
BabyVision (Chen et al., 2026a)	✓						✓	✓		✓		✓							✓		✓	
Spatial-DISE (Huang et al., 2025)				✓										✓		✓		✓				
MM-IQ (Cai et al., 2025a)					✓	✓	✓	✓		✓			✓								✓	✓
MMSI-Bench (Yang et al., 2025)				✓				✓				✓		✓		✓		✓				
TB-Bench (Charoenpitaks et al., 2025)		✓						✓				✓			✓							
RSIEval (Hu et al., 2023)			✓					✓	✓			✓						✓	✓			
GEOBench-VLM (Danish et al., 2024)								✓	✓			✓						✓				
SpatiaLab (**?**)				✓								✓		✓		✓		✓				
MCBench (Xu et al., 2024)												✓		✓								
WearVQA (Chang et al., 2025)		✓	✓					✓	✓			✓						✓	✓			
MMMU (Yue et al., 2023)						✓			✓													
ZeroBench (Roberts et al., 2025)										✓												
MM-Vet (Yu et al., 2023)			✓			✓		✓	✓			✓										
LEGO-Puzzles (Tang et al., 2025)				✓	✓		✓	✓					✓	✓							✓	
Q-Spatial-Bench (Cheng et al., 2024b)				✓												✓	✓					
CausalSpatial (Ma et al., 2026b)	✓			✓			✓					✓										
CRAG-MM (Wang et al., 2025a)		✓	✓						✓			✓						✓	✓			
EarthVQA (Wang et al., 2023)								✓	✓					✓								
MARVEL (Jiang et al., 2024)					✓		✓	✓		✓			✓								✓	✓
LiveVQA (Fu et al., 2025)		✓	✓						✓			✓							✓		✓	
VLM4Bio (Maruf et al., 2024)								✓	✓			✓						✓				
ScienceQA (Lu et al., 2022)						✓			✓													
SimpleVQA (Cheng et al., 2025)		✓	✓																			
DARE (Sterz et al., 2024)			✓						✓													✓
HAL Spatial (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 5. Comparison of multimodal and spatial reasoning benchmarks.

Benchmark	**Selected Tasks**	**Number of Images**	**Number of Total Tasks**	**Publish Date**
OmniBench	50	11525	11000	Sep-2024
Spatial457	781	1000	68752	Feb-2025
MathVision	2476	3472	3040	Feb-2024
BabyVision	82	388	388	Jan-2026
Spatial-DISE	301	559	559	Oct-2025
MM-IQ	145	2710	2710	Feb-2025
TB-Bench	938	2000	2000	Jan-2025
RSIEval	137	100	936	Jul-2023
GEOBench-VLM	61	–	10000	Nov-2024
SpatiaLab	617	1400	1400	Feb-2026
WearVQA	259	2520	2520	Nov-2025
MMMU	51	10500	10500	Nov-2023
zerobench	73	100	434	Feb-2025
MM-vet	55	200	218	Aug-2023
LEGO-Puzzles	580	1100	1100	Mar-2025
MMSI-Bench	623	1990	1000	May-2025
Q-Spatial-Bench	82	1406	1406	Jun-2024
CausalSpatial	322	–	1541	Jan-2026
crag-mm	8	8500	8500	Oct-2025
EarthVqa	50	6000	208593	Dec-2023
MARVEL	97	770	770	Apr-2024
DARE	50	–	5525	Sep-2024

Table 6. **Hierarchical task taxonomy of the HolisticSpatial benchmark.** Tasks are organised into three levels: (L1) domain categories capturing distinct spatial reasoning modalities, (L2) sub-categories reflecting specific cognitive or perceptual operations, and (L3) fine-grained task types with defined evaluation formats. Counts indicate the number of samples per task.

Category	Sub-Category	Task	Count
Physical Reasoning	Predictive Physics	Predictive	50
		Planning	135
	Navigation & Control	Turning – Car	82
		Navigation	177
		Lane Position	228
		Turning – Entity	214
		Lane Change	180
Positional & Geometric Reasoning	3D Reasoning	3D Pose	226
		3D Shape Finding	56
		3D Geometry	349
		3D Assembly	119
		3D Rotation	197
		3D Object Detection	151
		3D Folding	68
		3D Projection	50
	Geometric Reasoning	Occluded Attributes	374
		Occluder Attributes	181
		Direction and Orientation	145
		Compatibility	88
	2D Reasoning	2D Geometry	102
	Positional Relations	Region–Region	79
		Camera–Region	81
		Object–Region	57
		Camera–Object	86
		Object–Object	87
	Localization	Relative Location	50
		Object Localization	50
Knowledge-Grounded Reasoning	Mathematical Reasoning	Arithmetic	107
		Algebra	299
		Scaling & Ratio	143
		Geometry – Transformational	157
		Geometry – Coordinate	77
		Geometry – Descriptive	98
		Constraint Geometry	87
		Metric Geometry	272
		Shape Decomposition	205
		Angle Reasoning	143
	Creative & Abstract	Pattern	59
		Riddle	73
		Trick	50
	Domain Specific	Engineering & Technology	51
Relational & Causal Reasoning	Causal Reasoning	Logical Counterfactual	113
		Physical Causality	62
	Relations	Spatial Relationship	280
Temporal & Dynamic Reasoning	Temporal Reasoning	Temporal Movement	51
		Sequencing	47
	Motion Reasoning	Camera Motion	70
		Object Motion	50
Quantitative & Comparative Reasoning	Measurement	Relative Distance	234
		Measurement	61
		Absolute Distance	82
	Counting	Counting	76
Perceptual Understanding	Object-Level	Size & Scale	170
		Attribute	53
		Object Identification	50
		Color	50
	Scene-Level	Scene Understanding	50

B. Task Taxonomy and Descriptions

Understanding spatial reasoning requires more than simple object recognition; it depends on how objects relate, align, and interact within physical space. To capture this complexity, we introduce *HolisticSpatial*, a benchmark designed to evaluate fine-grained spatial reasoning across diverse settings. The benchmark covers a wide spectrum of visual spatial-reasoning abilities, organised into a three-level hierarchy.

At the *first level* (**L1**), we define eight mutually exclusive **domain categories** that capture distinct reasoning modalities. The *second level* (**L2**) partitions each domain by the specific **cognitive or perceptual operations** involved. The *third level* (**L3**) further refines these into fine-grained **task types** with well-defined question formats and evaluation criteria. Table 6 summarises the full hierarchy along with sample counts.

The tasks in *HolisticSpatial* are designed to reflect how humans naturally reason about environments, ranging from relative positioning (e.g., *What is to the left of the microwave?*) to geometric and physical inference (e.g., *Would this stack fall if another book is added?*). Each category is further decomposed into subcategories that capture critical dimensions of reasoning, such as occlusion, orientation, navigation, and physical interactions like gravity and collision.

By framing subtasks in intuitive, everyday terms, the benchmark encourages models to move beyond object detection toward structured spatial interpretation. This design enables systematic evaluation of generalization across contexts relevant to real-world applications, including robotics, augmented and virtual reality, autonomous driving, and human-computer interaction. The following sections provide concise task-level descriptions for each category and subcategory.

B.1. Physical Reasoning

Physical Reasoning tasks require a model to apply knowledge of real-world physical laws, kinematics, collision dynamics, navigation constraints, and spatial affordances, to interpret or predict states and events depicted in images. With $n = 1{,}312$ items, this domain spans three second-level categories that together cover predictive simulation, embodied navigation, and physical interaction.

B.1.1. PREDICTIVE PHYSICS

Predictive Physics tasks ask models to reason about what will happen next given the current physical configuration shown in an image, requiring an implicit forward simulation of physical principles rather than mere scene description. The $n = 185$ items in this category are divided between unconstrained single-outcome prediction and goal-directed multi-step planning.

Predictive. Predictive tasks ($n = 50$) present a static snapshot of a physical scene and ask the model to forecast the most probable next physical state without explicit velocity or force parameters. Given an image of a heavy hammock attached to the edge of an overhanging roof, the model must evaluate structural load and cantilever mechanics to judge the most likely outcome, testing whether it can simulate gravitational and mechanical cause-and-effect chains purely from a single observed frame. A complementary task provides a top-down diagram of billiard balls with a directional arrow on the cue ball and asks whether any ball will enter a pocket, exercising two-dimensional Newtonian trajectory reasoning from a visually encoded motion description.

Planning. Planning tasks ($n = 135$) extend pure prediction by requiring the model to identify a sequence of interventions or actions necessary to achieve a specified goal state while respecting physical constraints. Given a bird's-eye image of a toy car on a collision course with several objects, a stack of books, a vase, and a chair, the model must determine which specific objects need to be removed to guarantee a collision-free path, testing hierarchical physical planning under real-world spatial constraints. A second representative task presents a LEGO build with a defined target state and a set of candidate intermediate steps, asking the model to arrange those steps into the correct construction sequence, exercising procedural planning over a discrete, structured physical state space.

B.1.2. NAVIGATION & CONTROL

Navigation & Control tasks ground physical reasoning within road-driving scenarios, requiring models to parse the spatial configuration of lanes, vehicles, and traffic participants from egocentric dashcam images and to produce correct kinematic or positional judgements. This is the largest second-level category within Physical Reasoning ($n = 881$) and spans five sub-categories covering the ego vehicle's own motion as well as the motion and position of other traffic participants.

Turning – Car. Turning–Car tasks ($n = 82$) present a dashcam image and ask the model to describe or classify the ego-vehicle's current turning manoeuvre. Given a forward-facing dashcam frame captured mid-turn at an urban intersection, the model must determine whether the ego vehicle is turning left, turning right, or proceeding straight, requiring egocentric motion inference from road geometry, visible lane markings, and kerb curvature within a single static frame. These tasks specifically test whether a model can maintain a first-person spatial frame and distinguish self-motion cues from the surrounding scene context.

Navigation. Navigation tasks ($n = 177$) require the model to resolve directional or route-planning questions embedded within real or schematic map-like images. Given an overhead sign showing directional arrows and named destinations, the model must identify which road to take given a stated travel goal, testing the composition of cardinal-direction reasoning with symbolic map interpretation. A second example provides a multi-photograph room walk-through and asks in which direction a named landmark lies relative to the observer at the final step, exercising egocentric frame updating across a described movement sequence.

Lane Position. Lane Position tasks ($n = 228$) present a driving scene and ask the model to classify the current lane occupied by a specified traffic entity from a fixed option set: front lane, front-left lane, front-right lane, or oncoming-traffic lane. Given a dashcam image showing multiple vehicles ahead at varying lateral offsets, the model must identify the precise lane of a named entity by reading lane markings, vehicle silhouettes, and perspective cues, testing fine-grained spatial grounding of traffic participants within a structured road topology. Correct answers often require distinguishing positions that differ by a single lane-width, placing high demands on lateral localisation precision.

Turning – Entity. Turning–Entity tasks ($n = 214$) mirror the Turning–Car sub-category but direct the model's attention to other traffic participants, vehicles, cyclists, or pedestrians, rather than the ego vehicle. Given an image of a junction from the ego perspective, the model must identify whether a specified entity is turning left, turning right, merging, or proceeding straight, requiring allocentric tracking of another agent's motion trajectory within the shared scene. This tests whether the model can maintain an independent spatial reference frame for a non-ego agent while remaining anchored to the ego-vehicle viewpoint.

Lane Change. Lane Change tasks ($n = 180$) ask models to identify whether a named traffic entity is in the process of changing lanes and to characterise the direction and stage of the manoeuvre. Given a dashcam image capturing a vehicle straddling two lane markings, the model must infer the direction of movement from the vehicle's angle relative to the lane markings and its lateral offset from the lane centre, testing dynamic spatial inference from a single static frame. These items are particularly demanding because mid-manoeuvre positions are visually ambiguous and require integrating multiple weak spatial cues simultaneously.

B.1.3. PHYSICAL INTERACTION

Physical Interaction tasks require models to reason about what happens when objects are placed in proximity, when forces are applied, or when physical constraints between objects come into play, going beyond trajectory prediction to incorporate material, shape, and affordance reasoning. The $n = 246$ items are divided between property inference and collision analysis.

Property Inference. Property Inference tasks ($n = 74$) ask the model to deduce a physical property of an object, material, rigidity, weight, or spatial fit, that is not directly labelled in the image but can be inferred from visual appearance or geometric configuration. Given an image of a glass jar already containing an apple and a lemon, the model is asked whether a specified object positioned directly above the opening would fit inside if dropped, testing the ability to reason about relative size, container geometry, and gravity without any explicit measurements. A related task presents two utensils, a spoon and chopsticks, and asks whether they are made of the same material, exercising cross-object surface-texture and reflectance comparison under naturalistic viewing conditions.

Collision. Collision tasks ($n = 172$) require the model to predict whether a moving object will intersect with one or more obstacles and to determine what spatial interventions would prevent each collision. Given a bird's-eye image of a floor plan with a toy vehicle at a known position and heading, the model must trace the forward path, identify all objects in that path, and determine whether removing a single named obstacle is sufficient to clear the route, testing multi-obstacle spatial planning under a physical motion constraint. A second task type presents a soccer ball with a directional trajectory arrow

aimed at a goal and asks whether the ball will score, requiring 2D path-intersection reasoning between a directed vector and a bounded target region.

B.2. Geometric and Positional Reasoning

Geometric and Positional Reasoning is the largest domain in the benchmark ($n = 2{,}539$), covering the full range of spatial reasoning operations that act on explicit geometric structure: three-dimensional object manipulation, planar geometry, occlusion-based attribute inference, and precise positional relationship identification between scene elements. Its five second-level categories span from volumetric 3D reasoning down to two-dimensional planar geometry and fine-grained object localisation.

B.2.1. 3D REASONING

3D Reasoning tasks require models to interpret, manipulate, or reason about objects in three-dimensional space from one or more 2D image projections, covering mental rotation, depth-order inference, surface reconstruction, and structural assembly. With $n = 1{,}216$ items across eight sub-categories, this is the largest single second-level category in the benchmark and constitutes the primary evaluation axis for volumetric spatial cognition.

3D Pose. 3D Pose tasks ($n = 226$) present an object rendered from one viewpoint and ask the model to identify which of several candidate images shows the same object from a different, specified perspective. Given a LEGO model photographed from the front alongside four candidates showing it from the left, right, back, and above, the model must select the image corresponding to the named viewpoint, testing the ability to mentally traverse viewpoint space while preserving structural object identity. These tasks specifically probe whether a model maintains a stable allocentric 3D representation rather than relying on surface-texture recognition.

3D Shape Finding. 3D Shape Finding tasks ($n = 56$) provide a 2D cross-section, silhouette, or set of orthographic projections and ask the model to identify the corresponding 3D solid from a set of candidates. Given three orthographic views (top, front, side) of an object, the model must reconstruct the implicit 3D shape and select the matching solid from four rendered options, exercising shape-from-multiple-views reconstruction reasoning. The relatively small sample count reflects the high difficulty of sourcing unambiguous controlled multi-view stimuli with verified unique 3D solutions.

3D Geometry. 3D Geometry tasks ($n = 349$) cover questions about geometric properties of three-dimensional objects and scenes, including surface area, volume, dihedral angles, face adjacency, and symmetry axes. Given an isometric render of a compound solid composed of unit cubes, the model may be asked to count visible faces or determine the total surface area, requiring simultaneous tracking of face visibility and inter-cube occlusion from a fixed viewpoint. A related variant presents two highlighted components in a 3D scene and asks whether they share a face or are merely adjacent by a vertex, testing precision contact-geometry reasoning from a rendered perspective.

3D Assembly. 3D Assembly tasks ($n = 119$) present a target LEGO structure and an in-progress build state, and ask the model to predict the outcome of adding a specified piece, identify the correct attachment site, or judge whether the piece requires rotation before attachment. Given the current build state (x_1), the next brick (x_2), and the target final object (x_3), the model must select which of five candidates correctly shows the post-addition state, testing forward simulation of incremental 3D assembly under structural connector constraints. These items demand simultaneous maintenance of the current structural inventory, the new piece's geometry, and the valid attachment topology, a three-way compositional spatial reasoning requirement.

3D Rotation. 3D Rotation tasks ($n = 197$) ask the model to identify the clockwise rotation applied to a 3D object between a reference image and one or more rotated alternatives, or to select the image that corresponds to a specified rotation angle. Given a LEGO model in a top-down reference view (x_0) and four candidates rotated by $30°$, $60°$, $90°$, and $120°$ respectively, the model must identify the image matching the specified target rotation, requiring fine-grained quantitative angular discrimination from rendered 3D images. Unlike standard mental rotation paradigms, these tasks use discrete angle options, enabling direct tests of sensitivity to small angular differences (e.g., $30°$ vs. $60°$) that alter local feature positions but preserve global object identity.

3D Object Detection. 3D Object Detection tasks ($n = 151$) require the model to detect, count, or localise objects within a three-dimensional scene rendered as a static image, going beyond 2D detection to account for depth ordering and inter-object occlusion. Given an isometric LEGO scene with two bricks indicated by red arrows, the model must determine whether they are directly touching or separated, requiring stud-level contact detection within a cluttered 3D environment. A second variant provides a bird's-eye driving scene and asks for the count of a specific vehicle type, testing object enumeration under perspective foreshortening and partial occlusion.

3D Folding. 3D Folding tasks ($n = 68$) present a flat 2D net and ask the model to identify which 3D solid can be formed by folding it, or to determine which edge of the net coincides with a labelled reference edge after folding. Given a hexomino net representing a cube with uniquely coloured faces, four candidate cubes are shown with different face-colour arrangements, and the model must select the one consistent with the net's face adjacency layout, testing a complete two-dimensional-to-three-dimensional mental fold transformation. More complex variants use nets for octahedra or irregular polyhedra, requiring the model to generalise the folding operation beyond the most common training stimuli.

3D Projection. 3D Projection tasks ($n = 50$) present a 3D object with a directional arrow indicating a viewing axis and ask the model to select the corresponding 2D projection from a set of options. Given an isometric rendering of a multi-block solid with an arrow pointing along the depth axis, the model must identify the correct orthographic top-down view, exercising the inverse of 3D-pose reasoning: computing the 2D projection from a known 3D object and viewpoint rather than inferring the viewpoint from paired images. These tasks require the full projection pipeline, occlusion determination, foreshortening, and boundary computation, to be applied to a single static input.

B.2.2. GEOMETRIC REASONING

Geometric Reasoning tasks combine spatial perception with knowledge of geometric relationships, asking models to reason about occluded parts, directional alignments, and structural compatibility from partially informative visual inputs. The $n = 788$ items cover four complementary aspects of geometry-grounded perception.

Occluded Attributes. Occluded Attributes tasks ($n = 374$) present a scene in which one object is partially or fully hidden by another, and ask the model to infer a property of the hidden target object from contextual or associative visual cues. Given an image where a gray bicycle partially occludes a second object, the model must infer the colour or shape of the hidden object based on visible portions, the occlusion boundary, or prior knowledge of object co-occurrence patterns, testing inference under partial observability. These items evaluate whether models can maintain a probabilistic belief about unobserved scene regions using spatial reasoning and world knowledge rather than defaulting to a "cannot determine" response.

Occluder Attributes. Occluder Attributes tasks ($n = 181$) reverse the attribution direction: the object whose properties are unknown is the one doing the occluding, not the one being hidden. Given a scene where the occluding object's surface is in shadow or extends off- frame, the model must infer its shape or category from the silhouette of the occlusion boundary, the visible object edges, or contextual layout clues, testing shadow-geometry and silhouette-shape reasoning. This direction of spatial deduction is a less common evaluation paradigm and complements the more widely studied occluded-attribute inference task.

Direction and Orientation. Direction and Orientation tasks ($n = 145$) provide images containing spatially arranged objects or signs with explicit directional indicators, and ask the model to determine the orientation or facing direction of a specific element relative to a named reference. Given a street scene with one one-way sign facing south and a second sign nearby, the model must determine the cardinal direction the second sign faces, requiring correct allocentric direction composition in the presence of potential egocentric ambiguity. A complementary task shows a building photographed from the north and asks which side of the same building is shown in a second photograph, exercising viewpoint-based compass-direction inference.

Compatibility. Compatibility tasks ($n = 88$) ask whether two or more objects, shapes, or spatial configurations are geometrically compatible, whether they can be combined, fitted together, or assembled without violating geometric or physical constraints. Given a partial 2D shape alongside several candidate completion pieces, the model must identify which piece can be attached to form a specified target shape without overlap or gap, testing boundary-matching and shape-complementarity reasoning. These tasks require the model to mentally manipulate candidate pieces (including

possible rotation) and compare resulting contours to a target silhouette, a process analogous to jigsaw-piece fitting.

B.2.3. 2D REASONING

2D Reasoning tasks operate entirely in the image plane, asking models to reason about flat geometric shapes, planar relationships, and two-dimensional spatial configurations without any requirement for depth inference or viewpoint transformation. All $n = 102$ items fall within a single sub-category.

2D Geometry. 2D Geometry tasks ($n = 102$) present flat geometric figures, diagrams, or spatial patterns and ask quantitative or relational questions about their properties: side lengths, angles, areas, symmetry axes, or spatial arrangements among planar elements. Given a flat irregular figure alongside a dotted fold axis, the model must identify which of four candidate shapes results from folding one portion onto the other, testing planar fold transformation reasoning without any three-dimensional component. A second variant presents a matrix of geometric figures with one cell blank and asks the model to complete the pattern by selecting the missing shape, exercising rule-based 2D spatial sequence reasoning over rows and columns simultaneously.

B.2.4. POSITIONAL RELATIONS

Positional Relations tasks require models to identify or describe the precise spatial relationship between two specified scene elements, left/right, above/below, near/far, relative to a clearly stated reference. The $n = 333$ items are partitioned by the pair of entity types whose relationship is being queried, spanning region-to-region, camera-to-region, camera-to-object, and object-to-object configurations.

Positional Relationship (Region – Region). Region–Region tasks ($n = 79$) ask the model to characterise the spatial relationship between two defined scene regions, such as two rooms in a floor plan or two areas in an aerial map. Given a floor-plan diagram, the model may be asked which room lies to the west of a named corridor when the agent is facing south, requiring simultaneous map-reading, reference-frame establishment, and directional relation extraction from a schematic spatial layout. These tasks test allocentric spatial encoding over region-level scene elements rather than individual point objects.

Positional Relationship (Camera – Region). Camera–Region tasks ($n = 81$) ask the model to specify the position of a scene region relative to the camera's viewpoint at the moment of capture. Given a photograph of a billboard and the instruction that the photographer was facing north, the model must determine whether the billboard was to the photographer's left, right, in front of, or behind them, exercising egocentric direction binding between a camera pose and a named scene region. The added complexity relative to Region–Region tasks lies in adopting the camera as a moving, oriented reference entity rather than an absolute spatial anchor.

Positional Relationship (Object – Region). Object–Region tasks ($n = 57$) require the model to determine the spatial relationship between a named object and a broader scene region within a single image. Given an indoor or outdoor scene, the model must identify where a specific object is located relative to a defined region (e.g., left side of the room, upper area of the frame, or background zone). These tasks evaluate the ability to integrate object localisation with coarse regional segmentation, ensuring correct assignment of spatial position under an allocentric frame of reference.

Positional Relationship (Camera – Object). Camera–Object tasks ($n = 86$) narrow the spatial query from a region to a single object, asking the model to determine the object's position relative to the camera frame at a specific moment. Given two consecutive photographs taken during a room walk-through, the model must determine in which direction a specific object (e.g., the stove) lay relative to the photographer at the moment of the second image, requiring dynamic egocentric frame-of-reference updating across a described movement sequence. These tasks combine navigation and object-localisation reasoning within a single egocentric spatial query.

Positional Relationship (Object – Object). Object–Object tasks ($n = 87$) ask the model to express the spatial relationship between two explicitly named objects within a single image, determining whether one is in front of, behind, beside, or above the other. Given an indoor scene with multiple items, the model must identify where a named decorative element lies relative to the main table, testing direct allocentric object-to-object position extraction from a natural photograph. These

items require both accurate object localisation and the correct assignment of a directional preposition under a specified reference convention.

B.2.5. LOCALIZATION

Localization tasks require models to determine the position of an object or scene region in absolute or relative spatial terms, either producing a self-contained location description or selecting from a set of candidate positions. The $n = 100$ items are split equally between relative and absolute localisation paradigms.

Relative Location. Relative Location tasks ($n = 50$) ask the model to express the location of a target entity in terms of its relationship to a named anchor entity rather than in absolute image coordinates. Given a scene image and a specified target object, the model must describe where the target lies with respect to a named reference (e.g., above and to the left of the table), exercising relational spatial language production tied to a verified visual spatial configuration. These tasks evaluate the correspondence between the model's internal spatial representation and its capacity to translate that representation into standard relational directional language.

Object Localization. Object Localization tasks ($n = 50$) ask the model to confirm the presence of a named object in an image and, if present, to identify its approximate spatial location within the scene. Given an aerial or indoor image, the model must verify the existence of a named element and describe its position within the scene layout (e.g., foreground centre, upper-left quadrant), testing joint detection and coarse spatial grounding without requiring bounding-box precision. The emphasis is on correct spatial-region assignment rather than pixel-level localisation.

B.3. Knowledge-Grounded Reasoning

Knowledge-Grounded Reasoning tasks require models to apply formal domain knowledge, mathematics, abstract logic, or specialised discipline-specific concepts, in conjunction with visual perception to answer questions that cannot be resolved by spatial intuition alone. With $n = 2{,}332$ items, this is the second-largest domain, and its three second-level categories span formal mathematics, abstract pattern and lateral reasoning, and applied domain-specific knowledge.

B.3.1. MATHEMATICAL REASONING

Mathematical Reasoning tasks present visually rendered mathematical problems and require models to apply formal procedures to reach a numerical or symbolic answer. With $n = 2{,}049$ items across thirteen sub-categories, this is the most internally diverse category in the benchmark and constitutes the primary evaluation axis for quantitative visual reasoning.

Math – Arithmetic. Arithmetic tasks ($n = 107$) present images containing quantities, clocks, timelines, or symbolic addition problems and ask models to perform basic arithmetic to derive a numerical answer. Given an image of a competition timeline specifying the end time and a duration in hours and minutes, the model must compute the start time, requiring modular arithmetic applied to a visually presented temporal problem. These tasks test whether models can extract numerical premises from visual representations such as clocks or schedules and combine them through arithmetic without defaulting to language-prior heuristics.

Geometry – Transformational. Transformational Geometry tasks ($n = 157$) ask models to predict the outcome of a defined geometric transformation, reflection, rotation, translation, or fold, applied to a 2D figure. Given a flat shape alongside a dotted fold axis, the model must identify which of four candidates correctly depicts the result after folding, testing the ability to apply a single formal planar transformation to a novel geometric input. These tasks evaluate whether the model reasons about the transformation as a formal operation or merely recognises results by visual similarity.

Math – Algebra. Algebra tasks ($n = 299$) embed symbolic constraint systems in images, puzzle grids, letter-substitution sums, or visual equation sets, and ask the model to derive one or more unknown values. Given an image of an addition sum in which three letters each represent a distinct digit, the model must determine the sum of those three letters, requiring the formulation and solution of a constraint-satisfaction problem from visual premises. These items combine visual parsing of symbolic notation with algebraic constraint propagation, placing simultaneous demands on both modalities.

Scaling & Ratio. Scaling & Ratio tasks ($n = 143$) present images where two or more quantities are related by a proportional relationship, and ask the model to determine an unknown by applying ratio or proportion reasoning. Given two floor-plan images alongside a scale bar, the model must estimate the real-world dimension of a specified architectural element, requiring visual ratio extraction and its application to a known reference scale. A second variant shows two objects of different sizes and asks how many of the smaller would fit inside the larger, exercising container-size ratio reasoning.

Geometry – Coordinate. Coordinate Geometry tasks ($n = 77$) present labelled Cartesian coordinate diagrams or function graphs and ask the model to extract, compare, or reason about values at specific coordinates. Given a graph of a mathematical function, the model must identify the x-value at which the function achieves a specified y-value, requiring accurate visual axis-reading combined with function-value inference. These tasks test the intersection of coordinate-axis parsing precision and formal function-interpretation reasoning.

Geometry – Descriptive. Descriptive Geometry tasks ($n = 98$) ask the model to identify or describe formal geometric properties of named shapes, faces, edges, vertices, angles, symmetry, without numerical calculation. Given a 3D polyhedron image with labelled components, the model must identify which net edge coincides with a labelled reference after folding, testing structural component-identity tracking within a descriptive geometric framework. A second variant presents a geometric figure and asks the model to name the type or characterise a property of a specific sub-region, exercising label-to-shape correspondence in formal geometry.

Geometric Reasoning & Visual Problem Solving. This sub-category ($n = 234$) combines pattern-based geometric inference with visual puzzle-solving, requiring models to apply formal geometric rules and perceptual grouping simultaneously to identify missing elements or complete incomplete configurations. Given a 3×3 matrix of geometric figures with one cell blank, the model must infer the row-and-column transformation rule and select the correct completion from candidate options, testing rule induction from visual instances combined with formal geometric transformation knowledge. These items function as a visual analogue to abstract reasoning tests and probe generalisation from a small observed instance set to a systematic rule.

Constraint Geometry. Constraint Geometry tasks ($n = 87$) present a set of geometric constraints such as congruence, parallelism, or fixed angles, and ask the model to determine whether a given configuration satisfies them or to identify the configuration that does. Given a diagram with labelled angles and a question about whether two lines are parallel, the model must apply the parallel-line angle theorems to the visually presented configuration, testing formal geometric inference in a visual proof context. A second variant asks the model to determine the minimum number of conditions required to uniquely specify a geometric configuration, testing constraint- counting meta-reasoning.

Geometric Construction & Optimization. Construction & Optimization tasks ($n = 95$) ask the model to identify the result of a geometric construction procedure or to determine the configuration satisfying a geometric optimisation criterion. Given a diagram annotated with compass-and-straightedge construction steps, the model must identify the resulting construction, a perpendicular bisector, angle bisector, or circumscribed circle, testing knowledge of canonical construction procedures and their visual signatures. These items require recognising that a specific sequence of operations produces a uniquely identifiable geometric entity, going beyond simple shape recognition.

Metric Geometry. Metric Geometry tasks ($n = 272$) require quantitative reasoning about geometric measurements: lengths, areas, perimeters, volumes, and distances between geometric entities in visually presented diagrams or scenes. Given a floor-plan image with scale markings, the model must compute the total perimeter of a specified room in real-world units, requiring scale extraction, unit conversion, and perimeter computation from an architectural layout. A second task type provides a coordinate plane with two marked points and asks the model to compute their Euclidean distance, testing metric formula application to visually presented spatial data.

Shape Decomposition. Shape Decomposition tasks ($n = 205$) present a complex figure and ask the model to identify which simpler component shapes it can be decomposed into, or to identify the compound shape formed by combining two or more given primitives. Given an image of an irregular polygon, the model must determine which combination of rectangles and triangles tiles the original shape without overlap or gap, testing compositional geometric reasoning and area-conservation awareness. These tasks probe whether models represent compound shapes as structured compositions of parts rather than as holistic visual templates.

Circle Geometry. Circle Geometry tasks ($n = 132$) involve circles, arcs, chords, tangents, and inscribed or circumscribed polygons, asking models to compute or reason about angle, arc length, area, or relationship properties. Given a diagram with a circle, a labelled diameter, and an inscribed angle, the model must apply the inscribed-angle theorem to determine the angle's measure, testing knowledge of a specific circle-geometry theorem in a visual application context. A second variant shows two tangent circles with a shared external tangent and asks for the length of the tangent segment, requiring the combination of construction knowledge with metric computation.

Angle Reasoning. Angle Reasoning tasks ($n = 143$) focus specifically on inferring, computing, or comparing angles in static images of geometric configurations or real-world scenes. Given an image of a leaning building with a visible reference vertical, the model must estimate the angle of inclination from the vertical in degrees, testing angle-estimation reasoning applied to a real-world visual context without explicit measurement tools. A second variant provides a geometric figure with several labelled angles and asks the model to derive an unlabelled angle by applying angle-sum or supplementary-angle rules, testing formal angle computation in a diagram-reading context.

B.3.2. CREATIVE & ABSTRACT KNOWLEDGE

Creative & Abstract Knowledge tasks require models to apply non-formal or divergent reasoning strategies, pattern induction, lateral thinking, and misdirection detection, in contexts where standard algorithmic approaches are insufficient or deliberately misleading. The $n = 182$ items span three sub-categories that collectively test flexible and adaptive reasoning from visual premises.

Pattern. Pattern tasks ($n = 59$) present a sequence or matrix of visual stimuli and ask the model to identify the underlying rule and use it to complete or continue the sequence. Given a set of six geometric figures and an instruction to divide them into two groups each sharing a common characteristic, the model must induce the classification criterion from the visual examples alone, testing unsupervised rule extraction from a small, visually presented instance set. These items go beyond simple perceptual grouping to require identifying an abstract structural invariant that holds across every member of each category.

Riddle. Riddle tasks ($n = 73$) embed a logical deduction or elimination puzzle in image form and ask the model to derive the unique correct answer by systematically ruling out inconsistent possibilities. Given a set of five measuring jugs filled to different levels with the stated constraint that four contain identical amounts and one is different, the model must compare the visible water levels to identify the odd jug, requiring quantitative visual comparison and elimination-based logical reasoning. These tasks test whether models apply systematic deduction to visually presented constraint sets rather than guessing by visual salience.

Trick. Trick tasks ($n = 50$) present problems that contain deliberate visual or logical misdirection, where correct answers require resisting the obvious interpretation and identifying the non-obvious solution. Given a narrative-style puzzle with a stated arithmetic component alongside an image that encodes the answer as a visual cue, the model must recognise that the relevant information is embedded visually rather than derivable arithmetically, testing metacognitive awareness about when visual evidence should override computational strategies. These tasks evaluate a form of epistemic flexibility, knowing which reasoning mode to activate, that is not captured by standard mathematical or perceptual evaluation paradigms.

B.3.3. DOMAIN SPECIFIC

Domain Specific tasks require models to apply specialised knowledge from engineering, technology, or natural science to interpret visual content that cannot be correctly answered by general spatial or mathematical reasoning alone. The $n = 101$ items are split between two applied knowledge domains.

Engineering & Technology. Engineering & Technology tasks ($n = 51$) present images of technical systems, circuit diagrams, logic-gate schematics, or mechanical assemblies, and ask the model to interpret their function, identify components, or reason about their operational behaviour. Given an image of a logic-gate circuit with labelled inputs, the model must determine the output signal for a specified input configuration, requiring both visual parsing of the gate schematic and application of Boolean logic rules. These items sit at the intersection of visual diagram-reading and formal domain-knowledge application, evaluating whether models have acquired functional knowledge of standard technical representations.

Scientific Reasoning. Scientific Reasoning tasks ($n = 50$) present images from natural or applied science contexts, medical scans, biological process diagrams, or physical experiment setups, and ask the model to apply domain knowledge to interpret the visual content. Given a chest X-ray image with visible anatomical structures, the model must identify which, if any, of a named list of pathological conditions are present, requiring the application of radiological knowledge to a clinically realistic visual input. A second task type provides a biological schematic depicting nutrient transport pathways and asks where a specific substance is absorbed, testing knowledge- grounded causal reasoning applied to a scientific diagram.

B.4. Relational, Logical & Causal Reasoning

Relational, Logical & Causal Reasoning tasks require models to reason about relationships between entities, whether causal, counterfactual, or purely spatial, going beyond direct perception to demand inferential reasoning about how entities interact or depend on one another. With $n = 455$ items, this domain spans two second-level categories that together evaluate causal inference and explicit spatial relation identification.

B.4.1. CAUSAL REASONING

Causal Reasoning tasks require the model to reason about cause-and-effect relationships depicted or implied in images, either by predicting what will follow from an observed state or by evaluating counterfactual scenarios in which an aspect of the situation is changed. The $n = 175$ items are divided between counterfactual and physical causality paradigms.

Logical Counterfactual. Logical Counterfactual tasks ($n = 113$) present a scene and ask what would happen if a specific element were removed, changed, or absent, requiring the model to construct a counterfactual world model and reason about its consequences. Given a road scene with a suggested connecting path between two lanes, the question asks whether such a path truly exists under a counterfactual reading of the scene's geometry, testing whether the model can distinguish factual from hypothetical spatial queries. These tasks evaluate whether models maintain a distinct counterfactual reasoning mode rather than merely describing what is visually present.

Physical Causality. Physical Causality tasks ($n = 62$) ask the model to reason about direct physical cause-and-effect relationships that are visible in or inferable from the image. Given an image of a person attempting to hang a very heavy hammock from the overhanging edge of a roof, the model must determine the most likely physical outcome by reasoning about structural load, cantilever mechanics, and material limits, exercising qualitative physics applied to a naturalistic visual scene. These tasks require the model to function as an approximate physical simulator, predicting outcomes from observed structural configurations without access to numerical physics parameters.

B.4.2. RELATIONS

Relations tasks ask models to identify or describe a spatial relationship between two specified entities, testing the core capacity to encode directional, positional, and topological relations as explicit answers. All $n = 280$ items fall within a single sub-category.

Spatial Relationship. Spatial Relationship tasks ($n = 280$) present a scene and ask the model to characterise the spatial relationship between two named entities using directional or relational language. Given an indoor scene with a table and a chair, the model must determine and express whether the chair is in front of, behind, to the left of, or to the right of the table from the camera's perspective, testing precise directional relation extraction between two objects under a specified reference frame. These items represent the most direct evaluation of relational spatial language grounding and serve as a core probe of object-to-object spatial binding capacity.

B.5. Temporal & Dynamic Reasoning

Temporal & Dynamic Reasoning tasks require models to interpret motion, change, or time-ordered sequences, either from multi-image series or from temporal cues embedded in a single frame such as motion arrows, trajectory lines, or described past and future states. With $n = 218$ items, this domain covers temporal sequencing and motion inference across two second-level categories.

B.5.1. TEMPORAL REASONING

Temporal Reasoning tasks require models to process multi-image sequences or described time series and to reason about temporal order, relative timing, or causal succession of depicted events. The $n = 98$ items are split between movement tracking across frames and discrete event sequencing.

Temporal Movement. Temporal Movement tasks ($n = 51$) present a series of images representing consecutive moments and ask the model to determine the position, direction, or relationship of an entity at a specific time step, requiring temporal tracking from sequential visual evidence. Given a series of first-person photographs taken while walking through a building, the model is asked to identify the direction of a specific landmark relative to the observer at the moment of the final photograph, requiring integration of successive viewpoint changes across the image sequence. These tasks test dynamic egocentric frame-of-reference maintenance across a multi-step described movement, a capability distinct from single-frame spatial grounding.

Sequencing. Sequencing tasks ($n = 47$) present a set of images depicting different stages of a process and ask the model to arrange them in correct chronological or procedural order. Given four images showing intermediate LEGO assembly states between a defined start and a target, the model must arrange them into the correct build sequence, requiring simultaneous 3D state-matching and procedural temporal ordering. These items probe whether models have an implicit sense of process directionality and can distinguish valid from invalid orderings of incremental state transitions.

B.5.2. MOTION REASONING

Motion Reasoning tasks require models to infer motion properties, direction, trajectory, or relative displacement, from static visual cues such as optical flow indicators, motion arrows, blurred contours, or geometric trajectory markers embedded in individual images. The $n = 120$ items divide between camera motion inference and object motion inference.

Camera Motion. Camera Motion tasks ($n = 70$) provide a pair of images taken from the same scene and ask the model to infer the direction or type of camera movement that occurred between them. Given two photographs of the same room taken in succession, the model must determine from which direction the second photograph was taken relative to the first, inferring camera rotation or translation from the shift in framing, perspective, and the set of visible objects. These tasks require the model to distinguish camera motion from object motion and to maintain a consistent scene geometry across two frames.

Object Motion. Object Motion tasks ($n = 50$) ask the model to determine the direction or trajectory of a moving object depicted in a single static image, using embedded cues such as directional arrows, trajectory lines, or geometric configuration. Given a top-down image of a soccer ball with a directional arrow indicating its path, the model must determine whether the trajectory will cause the ball to enter a goal, requiring 2D path-intersection reasoning between the arrow vector and the spatial boundary of the target region. These tasks test whether models can interpret formal motion representations embedded in otherwise static images and use them for spatial forward prediction.

B.6. Quantitative & Comparative Reasoning

Quantitative & Comparative Reasoning tasks require models to produce or reason with numerical values derived from spatial or visual properties of scenes, including metric distances, relative size comparisons, and discrete object counts. With $n = 503$ items across two second-level categories, this domain assesses numerical spatial grounding spanning estimation, measurement, and enumeration.

B.6.1. MEASUREMENT

Measurement tasks require the model to estimate, compare, or compute metric or relative spatial quantities from visual evidence, such as the distance between two objects, the height of a structure, or the extent of a region. The $n = 377$ items are distributed across three sub-categories spanning relative, unit-based, and absolute measurement.

Relative Distance. Relative Distance tasks ($n = 234$) ask the model to compare the distances between pairs of objects and determine which is closer, farther, or proportionally related to a reference length. Given a dashcam image with multiple vehicles and a crosswalk visible ahead, the model must determine how many taxi-cab lengths fit between a building entrance and the crosswalk, requiring proportional spatial estimation from a single perspective-projected frame. These items test

whether models can extract and compare distance ratios from 2D projective images without explicit depth or camera calibration data.

Measurement. Unit Measurement tasks ($n = 61$) ask the model to estimate or compute a physical quantity expressed in standard units, length in metres, inclination in degrees, or area in square metres, from visual cues or scale markers. Given an image of a leaning structure with a visible reference object of known height, the model must estimate the structure's angle of inclination in degrees, requiring visual angle-estimation from a naturalistic scene without a protractor or explicit angular indicator. These tasks test calibrated visual quantification, requiring models to move beyond ordinal comparisons to produce precise numerical estimates grounded in a standard unit system.

Absolute Distance. Absolute Distance tasks ($n = 82$) require the model to estimate the metric distance between two specific entities in real-world units, typically in metres between traffic participants captured in driving scenes. Given a dashcam frame showing the ego vehicle and a named target entity, the model must estimate the straight-line distance in metres between them, requiring an implicit mapping from pixel-space offsets to calibrated real-world metres using learned scene-scale priors. These tasks are among the most challenging quantitative items in the benchmark, as they demand absolute metric calibration from uncalibrated monocular images.

B.6.2. COUNTING

Counting tasks require models to enumerate discrete objects or category instances within images, either through direct enumeration or by applying a relational filter that restricts the counted set. The $n = 126$ items split between straightforward category-level counting and relation-conditioned counting.

Counting. Direct Counting tasks ($n = 76$) ask the model to count all visible instances of a named object category in the image without any relational qualifier. Given a photograph of a cluttered table with multiple beer bottles at varying degrees of occlusion, the model must produce the exact total count, testing exhaustive visual enumeration under partial occlusion and scene clutter. A second task type presents a building exterior and asks how many chimneys are visible, exercising category-specific object enumeration in a structured outdoor scene.

Relational Counting. Relational Counting tasks ($n = 50$) add a spatial or state qualifier to the counting question, restricting the counted set to objects satisfying a defined relational criterion. Given an image of a bar scene with multiple glasses in various states of fullness, the model must count only the empty glasses that are positioned to the right of a named person, requiring the model to jointly apply a spatial filter and an attribute filter before enumeration. These tasks test whether models can compose relational predicates with counting operations, a substantially more demanding form of enumeration than simple category-based object counting.

B.7. Perceptual Understanding

Perceptual Understanding tasks require models to correctly identify, classify, or describe the visual properties of objects or scenes at the level of direct sensory perception, without requiring spatial transformation, causal inference, or formal mathematical reasoning. With $n = 422$ items, this domain covers both fine-grained object attribute recognition and holistic scene-level classification.

B.7.1. OBJECT-LEVEL PERCEPTION

Object-Level Perception tasks focus on the intrinsic visual properties of individual objects, size, colour, texture, shape, and general appearance, as directly observable in the image without requiring spatial reasoning beyond basic object identification. The $n = 372$ items span five attribute-specific sub-categories.

Size & Scale. Size & Scale tasks ($n = 170$) ask models to assess the relative or absolute size of named objects, either by comparing two objects within the same image or by estimating real-world dimensions from contextual scale cues. Given two photographs of different buildings placed side by side, the model must determine which building's first four floors are collectively taller, requiring cross-image size normalisation and comparative height reasoning. These tasks test whether models have calibrated size priors for familiar object categories and can apply them to make consistent cross-instance comparisons.

Attribute. Attribute tasks ($n = 53$) ask the model to identify a specific named property of an object, material, texture, surface finish, or structural category, from visual appearance alone. Given a photograph of a table setting with a spoon and chopsticks visible, the model must determine whether both utensils are made of the same material by comparing their surface reflectance, texture, and sheen, testing cross-object material-attribute comparison under naturalistic conditions. These tasks evaluate whether models have learned visually grounded attribute representations rather than relying on category-level material priors.

Exist or Object Identification. Exist or Object Identification tasks ($n = 50$) ask the model to confirm the presence of a specified object in the scene or to identify which of several candidate objects is depicted. Given an aerial image, the model must determine whether a named structural element is present and, if so, how many instances are visible, testing basic object detection and binary existence verification. These items form the perceptual foundation on which all higher-order spatial and causal reasoning tasks depend; failures here cascade into all other task categories.

Color. Color tasks ($n = 50$) ask the model to identify the colour of a specified object or region from a naturalistic image, requiring accurate colour perception under variable lighting conditions and potential colour-similarity confounds. Given a tractor image with a distinctively coloured mudguard component, the model must name the mudguard's colour precisely, testing colour identification under real-world lighting variation, partial occlusion, and the presence of nearby objects with similar hues. While perceptually simple in appearance, these tasks surface systematic weaknesses in models whose colour representations are dominated by category-level priors rather than instance-level image-patch analysis.

Appearance. Appearance tasks ($n = 49$) ask the model to describe or classify the general visual appearance of an object or scene in terms that go beyond a single named attribute, encompassing overall form, condition, or stylistic character. Given an image of a building's reflection in water, the model must characterise the perceptual appearance of the reflected image, whether distorted, inverted, rippled, or clear, testing holistic appearance description under a non-canonical viewing transformation. These tasks probe the model's capacity for qualitative perceptual description beyond categorical labelling, evaluating the richness and accuracy of its visual vocabulary.

B.7.2. SCENE-LEVEL PERCEPTION

Scene-Level Perception tasks require models to form a holistic understanding of the entire image as an integrated scene, classifying or describing its overall character rather than the properties of any individual object within it. All $n = 50$ items fall within a single sub-category.

Scene Understanding. Scene Understanding tasks ($n = 50$) present diverse images, aerial photographs, indoor environments, outdoor landscapes, or specialised domain scenes, and ask the model to classify the primary scene type or characterise the overall context. Given an aerial photograph, the model must determine whether the depicted environment is primarily residential, agricultural, industrial, natural, or urban, requiring integration of global layout cues such as road density, building footprints, and vegetation coverage that are distributed across the entire image rather than concentrated in any single object. These tasks test whether models can form stable holistic scene representations that go beyond the aggregate of individually identified objects, evaluating the top-down perceptual abstraction capacity underlying scene-level cognition.

As described, the taxonomy is designed as a hierarchical decomposition of spatial and multimodal reasoning into six primary domains that collectively span perceptual, geometric, logical, temporal, and knowledge-based cognition. At the highest level, each category isolates a distinct reasoning modality, while maintaining consistent structure to support cross-category comparison. Within each domain, L2 subcategories capture intermediate cognitive operations such as prediction, transformation, relation extraction, or measurement. This separation ensures that semantically similar tasks are grouped even when their surface form differs. L3 task types provide fine-grained operational definitions, making each unit evaluable with clear input-output expectations. The design balances coverage of low-level perceptual tasks with higher-level abstract reasoning, ensuring representation from raw perception to symbolic reasoning. Physical Reasoning explicitly models interaction, navigation, and dynamics, capturing embodied understanding of environments. Positional and Geometric Reasoning expands this with structured spatial inference across 2D, 3D, and relational settings, including occlusion and viewpoint transformations. Knowledge-Grounded Reasoning integrates mathematical, creative, and domain-specific knowledge to test structured reasoning beyond perception alone. Relational, Temporal, and Quantitative categories further isolate causal structure, time-dependent change, and numeric comparison, which are often conflated in prior benchmarks. Perceptual Understanding anchors the taxonomy at the sensory level, grounding higher reasoning in object and scene-level

recognition. Overall, the taxonomy is designed to be both exhaustive in coverage and modular in structure, enabling controlled evaluation of distinct reasoning capabilities while minimizing overlap between task families.

C. Automated Dataset Curation Pipeline

This appendix provides a complete technical specification of the four-stage automated curation pipeline used to construct FRONTIERSPATIAL from a raw pool of approximately 50,000 candidate instances drawn from 23 source datasets. The pipeline is designed as a *principled, reproducible methodology for holistic benchmark construction* — not merely an engineering description of how our dataset was built, but a transferable framework applicable to any multimodal reasoning domain. Each stage applies explicit, auditable quality criteria, logs the reason for every rejection, and reduces the pool at a different cost-quality tradeoff point. The final benchmark of 7,838 instances represents the intersection of structural validity, visual uniqueness, text-level quality, and multimodal coherence.

Figure **??** summarizes the full pipeline; Table 7 reports the instance counts and reduction at each stage.

Table 7. Instance counts and reductions at each pipeline stage.

Stage	Step	Remaining	Reduced By
Raw pool	—	50,000	—
Stage 1 Structural Heuristics	1a. Metadata integrity	44,402	5,598
	1b. Publication recency	43,852	550
	1c. Question length filter	35,232	8,620
	1d. Exact-text deduplication	26,683	8,549
	1e. Taxonomy remapping & pruning	24,830	1,853
Stage 2 Vision-Space Deduplication	2a+2b. Near-duplicate detection	15,792	9,088
	2c. Visual domain clustering	12,914	2,878
Stage 3 Text-Only LLM Scoring	R1. Spatial relevance (≥ 3)	8,814	4,100
	R2. Cognitive depth (≥ 2)	5,857	2,957
	R3. Answerability (≥ 3)	1,943	3,914
	R4. Language clarity (≥ 3)	9,812	4
Stage 4 Multimodal LLM Final Pass	R5. Visual necessity	9,411	405
	R6. Image-question coherence	8,912	499
	R7. Spatial signal quality	7,838	1,074
Final benchmark	—	**7,838**	**42,162**

The four stages operate at increasing computational cost and decreasing filter aggressiveness. Stages 1 and 2 are deterministic and LLM-free; Stage 3 uses text-only LLM scoring at low cost; Stage 4 applies full multimodal LLM evaluation to the surviving pool. This design ensures that expensive computation is reserved only for candidates that have already passed cheaper quality gates.

C.1. Stage 1: Structural Heuristics (60K+ $\rightarrow$ 24,830)

Stage 1 applies five deterministic, LLM-free filters in sequence. These filters are fast, cost-free, and eliminate structurally invalid or redundant instances that would otherwise waste LLM budget in downstream stages. Each filter logs the rejection reason per instance, making the pipeline fully auditable.

1a. Metadata integrity ($-$5,598). Any instance is discarded if (i) its associated image path is missing or does not resolve to a valid file on disk, (ii) for multiple-choice questions, the annotated answer is not a member of the provided choices list, or (iii) for open-ended questions, the answer field is empty or null. This filter removes broken instances silently inherited from source datasets — a pervasive issue when aggregating across heterogeneous annotation pipelines.

1b. Publication recency ($-$550). Instances originating from datasets published before 2020 are flagged for review. Items retained from pre-2020 datasets must belong to landmark collections with no modern equivalent; all others are removed. This filter eliminates annotation artifacts from older VQA practices, including inconsistent spatial labelling and outdated visual content.

1c. Question length filter ($-$8,620). A token-length window of $[10, 200]$ tokens is applied to all question strings. Questions below the lower bound (e.g., "*What is here?*") lack the specificity needed for spatial reasoning evaluation. Questions above the upper bound frequently embed the answer in an extended preamble or introduce text-only solvability — a shortcut that

circumvents visual inference. This single filter accounts for the largest reduction in Stage 1.

1d. Exact-text deduplication ($-$8,549). Each question string is lowercased, whitespace-stripped, and SHA-256 hashed. Instances sharing a hash across source datasets are treated as exact duplicates; only one copy is retained, preferring the instance with richer metadata or higher-resolution imagery. This step is particularly important given that several source datasets in FRONTIERSPATIAL share visual question sources (e.g., MMSI-Bench appears in the pool under multiple dataset identifiers).

1e. Taxonomy remapping and non-spatial pruning ($-$1,853). Each instance's source category label is mapped to FRONTIERSPATIAL four-level taxonomy (Section **??**). Instances that cannot be mapped to any of the 65 tasks in our schema after remapping are removed as non-spatial. This step constitutes the most intellectually substantive filter in Stage 1: it operationalizes the taxonomy as a quality gate, ensuring that every surviving instance is anchored to a defined spatial reasoning skill. For any new domain, only the remapping rules require redefinition — the remaining Stage 1 filters are domain-universal.

C.2. Stage 2: Vision-Space Deduplication (24,830 $\rightarrow$ 12,914)

Stage 2 addresses a limitation unaddressed by any prior spatial benchmark paper: when a benchmark is constructed from multiple source datasets, near-identical images frequently appear under different question phrasings. Text-level deduplication (Stage 1d) cannot detect these cases. Stage 2 operates entirely in visual embedding space.

2a. Image encoding. All unique images referenced across the surviving 24,830 instances are encoded using a CLIP ViT-L/14 encoder. For multi-image instances, a mean-pooled embedding is computed across the image list. All embeddings are stored in a FAISS flat inner-product index for efficient retrieval.

2b. Near-duplicate detection ($-$9,088). Cosine similarity is computed between all image embedding pairs. Instance pairs with similarity > 0.97 are treated as near-duplicates — corresponding to the same scene captured under minor viewpoint or lighting variation, or two datasets that independently cropped the same source image. Within each duplicate cluster, the instance with the higher composite text-quality signal (question length $\times$ answer specificity) is retained; all others are discarded. This step reveals that approximately 37% of the Stage 1 survivor pool contains near-duplicate visual content across source boundaries — a cross-dataset redundancy that would systematically inflate model performance if left uncorrected.

2c. Visual domain clustering ($-$2,878). K-means clustering ($K = 25$) is applied to all image embeddings to identify visual domain clusters (e.g., indoor scenes, outdoor scenes, aerial imagery, synthetic/abstract, medical). Any cluster comprising more than 15% of the post-deduplication pool is subject to stratified downsampling to enforce visual domain balance. This step prevents the benchmark from being dominated by a single visual context — a subtle failure mode of multi-source benchmarks that existing spatial evaluation suites have not explicitly addressed. The resulting pool of 12,914 instances exhibits a balanced visual domain distribution, verified by cluster entropy before proceeding to Stage 3.

C.3. Stage 3: Text-Only LLM Scoring (12,914 $\rightarrow$ 9,812)

Running 12,914 instances through a multimodal LLM would be prohibitively expensive. Stage 3 instead applies text-only LLM scoring — operating on the question, answer choices, category label, and metadata — at approximately one-tenth the cost of full multimodal evaluation. Each instance receives four scores in a single structured JSON output call, rated on a 1–4 ordinal scale.

R1. Spatial relevance (threshold ≥ 3, $-$4,100). *Does answering this question require spatial reasoning?*

1 Pure factual recall with no spatial component.

2 Spatial vocabulary present but answerable without spatial reasoning (e.g., "name an object to the left").

3 Requires interpreting spatial relationships from a scene.

4 Requires multi-step spatial reasoning (e.g., composing depth, occlusion, and orientation inferences).

This criterion filters questions that were tagged as spatial in source datasets but are answerable by language statistics alone — a documented failure mode of VQA benchmark construction (Goyal et al., 2017; **?**).

R2. Cognitive depth (threshold ≥ 2, $-2{,}957$). *What level of cognitive processing is required?*

1 Pure perception or object detection.

2 Spatial relation understanding.

3 Spatial transformation or inference.

4 Multi-step spatial planning or causal reasoning.

This score serves a dual purpose: as a quality gate (threshold ≥ 2) and as a difficulty signal for downstream taxonomic balancing. A benchmark composed exclusively of level-4 items is maximally hard but minimally diagnostic; maintaining a distribution across levels enables fine-grained model profiling.

R3. Answerability (threshold ≥ 3, $-3{,}914$). *Can this question be answered from the visual scene without external world knowledge?*

1 Requires domain-specific knowledge (medical, engineering).

2 Requires cultural or encyclopedic context.

3 Answerable from scene geometry and visual cues alone.

4 Purely spatial; no knowledge shortcut is available.

Items failing this criterion (score ≤ 2) are discarded entirely, as they conflate spatial reasoning with knowledge retrieval.

R4. Language clarity (threshold ≥ 3, -4). *Is the question unambiguously and grammatically phrased?*

1 Translation artifacts, encoding errors, or garbled text.

2 Ambiguous referents or underspecified spatial predicates.

3 Clear with minor phrasing issues.

4 Perfectly precise spatial language.

Instances scoring 1 or 2 are discarded. The near-zero reduction at this step (-4) reflects the effectiveness of Stages 1a–1d in eliminating malformed text, but the criterion remains essential as an explicit quality guarantee. An instance must pass all four thresholds simultaneously to proceed to Stage 4.

C.4. Stage 4: Multimodal LLM Final Pass (9,812 $\rightarrow$ 7,838)

Stage 4 applies the most expensive quality gate: full multimodal LLM evaluation. Having eliminated structural failures, visual duplicates, and text-level quality issues in prior stages, only 9,812 instances reach this point — making full multimodal scoring computationally feasible. Each instance is evaluated on three criteria requiring joint visual and linguistic reasoning.

R5. Visual necessity (threshold: discard if answerable without image, -405). The LLM is presented with the question and answer choices *without* the associated image. If the model selects the correct answer with high confidence from text alone, the instance is discarded as harbouring a language shortcut that circumvents visual spatial inference. This is the multimodal instantiation of R3 from Stage 3 — now empirically testable rather than heuristically estimated. It directly addresses the "language bias" problem documented across VQA benchmarks (Goyal et al., 2017; **?**; Kosoy et al., 2025).

R6. Image-question coherence (−499). The LLM is asked to verify that the image contains the spatial scene, objects, and relations referenced by the question. Instances where the LLM cannot identify the referenced entities in the image — indicating a misaligned question-image pair inherited from source datasets — are discarded. This filter is particularly important for instances drawn from general-purpose VQA corpora where spatial questions were not authored with their specific images in mind.

R7. Spatial signal quality (−1,074). The LLM scores the richness of spatial structure in the image on a 1–4 scale:

1 Flat, text-only, or isolated object with no scene context.

2 Minimal spatial structure (one foreground object, plain background).

3 Multi-object scene with discernible spatial relationships.

4 Rich spatial structure: depth cues, occlusion, perspective, or complex multi-object arrangements.

Instances scoring 1 or 2 are removed. This criterion serves as the final ranking signal within taxonomy cells: when a category exceeds its target quota, the lowest-scoring R7 instances are pruned first. The resulting 7,838-instance benchmark exhibits consistently high spatial signal quality across all 65 task types.

C.5. Generalizability of the Pipeline

The curation pipeline is designed to be domain-transferable. The structural heuristics of Stage 1 are universally applicable to any multimodal benchmark pool. The visual embedding operations of Stage 2 require no domain-specific configuration beyond the choice of encoder. The LLM scoring rubrics of Stages 3 and 4 are parameterized by the target reasoning taxonomy: to adapt the pipeline to a new domain (e.g., temporal reasoning, causal reasoning), only the R1 spatial relevance criterion and the taxonomy remapping rules of Step 1e require redefinition. All other filters transfer without modification.

We validate the pipeline's contribution empirically through a staged ablation: benchmark statistics and model performance are measured at each checkpoint (50K, 24K, 12K, 9K, 7.8K), demonstrating that each stage independently improves benchmark quality — reducing model score variance, increasing the human-model performance gap, and improving per-skill diagnostic specificity. These ablation results are reported in Appendix **??**.

Full code and running details of the pipeline is released and available in the submitted codes.

D. Automated Evaluation Pipeline

This appendix provides a complete technical specification of HARNESSEVALUATOR, the three-stage automated pipeline used to score, cluster, and statistically compare model outputs across all spatial reasoning benchmarks reported in the main paper. The pipeline is designed around three constraints: (i) *scalability* to hundreds of tasks and over a dozen models without human re-annotation at each run; (ii) *statistical rigour*, with all pairwise comparisons corrected for multiple testing and effect sizes reported alongside significance; and (iii) *resumability*, so that partial runs caused by API rate-limits or hardware interruptions do not require reprocessing already-scored rows. Section D.1 describes the overall architecture. Sections D.2–D.3 detail each stage. Section D.4 documents the output schema. Section D.5 discusses known limitations and design trade-offs.

D.1. Pipeline Overview

HARNESSEVALUATOR takes as input a root directory containing one sub-folder per evaluated model, each holding an `agent_response.csv` file with columns `id`, `task_description`, `answer`, and `response`. The pipeline then runs three sequential stages for every model folder and produces per-model and cross-model output artefacts:

Stage 1. Structured Signal Extraction. An LLM judge scores each response on four binary dimensions and writes results to `agent_scores.csv`.

Stage 2. Error Taxonomy Induction. Failed responses are embedded, reduced, clustered, and each cluster is assigned a taxonomy label from the E1–E7 spatial error vocabulary.

Stage 3. Cross-Model Diagnostic Synthesis. Bootstrap confidence intervals, McNemar tests (Bonferroni-corrected), and Cohen's d effect sizes are computed per model and per error type; results are exported as `model_evaluations.csv` and `model_evaluations.json`.

The entry point is the `HarnessEvaluator` constructor, which accepts four arguments summarised in Table 8.

Table 8. HARNESSEVALUATOR API parameters.

Parameter	Default	Description
`results_path`	*required*	Root directory; one sub-folder per model, each containing `agent_response.csv`.
`max_concurrent`	`10`	Thread-pool size for parallel LLM scoring in Stage 1.
`model_name`	`Gemini-3.1-Flash`	OpenAI-compatible model identifier for both scoring and cluster labeling.

D.2. Stage 1 — Structured Signal Extraction

Motivation. A single binary accuracy label is insufficient to localise *where* in the perception–grounding–reasoning pipeline a failure occurs. Stage 1 therefore decomposes each evaluation into four binary dimensions that collectively span the sub-processes required for correct spatial question answering.

Scoring dimensions. For each row in `agent_response.csv`, the LLM judge is prompted with the task description, ground-truth answer, and model response under a zero-shot structured prompt that requests a strict JSON object. Table 9 defines the four dimensions.

Table 9. Four binary scoring dimensions produced by Stage 1.

Dimension	Range	Definition
`accuracy_score`	$\{0, 1\}$	Final answer is correct (semantic equivalence for open-ended items).
`spatial_elements_score`	$\{0, 1\}$	Key spatial objects or entities are correctly identified.
`spatial_relations_score`	$\{0, 1\}$	Spatial relations (left/right/above/below/near/far) are correctly stated.
`spatial_reasoning_score`	$\{0, 1\}$	Logical spatial reasoning chain leading to the answer is correct.

Implementation. Scoring is parallelised over a `ThreadPoolExecutor` with `max_concurrent` workers; each worker submits a single chat-completion request and retries up to five times on transient failures before logging the row ID and returning the original unscored row. Incremental saving is performed after each completed future: the partial result frame

is written to disk immediately, so that a process crash does not lose already-completed scores. On re-runs, HARNESSEVALUATOR reads any existing `agent_scores.csv`, identifies rows with a non-null `accuracy_score`, and passes only the unscored rows to Stage 1, reducing redundant API calls.

Judge prompt design. The scoring prompt instructs the judge to return *only* a valid JSON object with no markdown fences, no trailing commas, and `null` (not `None`) for missing values. Temperature is fixed at 0.1 to minimise scoring variance; `max_tokens` is set to $1{,}000$ to accommodate the optional free-text `reasoning` field. For multiple-choice items, correctness is determined by exact string match after normalisation; for open-ended items, the judge applies semantic equivalence, which provides tolerance for paraphrase variants of a correct answer.

D.3. Stage 2 — Cross-Model Diagnostic Synthesis

Bootstrap confidence intervals. For each model, 500-resample bootstrap confidence intervals (95% CI) are computed over `accuracy_score` using the percentile method:

$$\widehat{\mu}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} \bar{X}_b, \qquad \text{CI} = \left[\hat{q}_{2.5},\ \hat{q}_{97.5}\right], \tag{1}$$

where $\bar{X}_b$ is the mean accuracy of the b-th bootstrap resample and $\hat{q}_p$ denotes the p-th empirical percentile over $\{\bar{X}_b\}$. NaN values are excluded before resampling; models with zero valid rows receive `mean = NaN`.

Pairwise McNemar test. All $\binom{N}{2}$ model pairs are compared on `accuracy_score` using the continuity-corrected McNemar statistic (**?**):

$$\chi^2_{\text{Mc}} = \frac{(|b-c|-1)^2}{b+c}, \tag{2}$$

where b (resp. c) counts instances on which model 1 succeeded and model 2 failed (resp. the reverse), computed on the intersection of task IDs shared by both models. The p-value is obtained from the $\chi^2(1)$ distribution. If $b + c = 0$ (models agree on all items), the statistic is set to 0 and $p = 1$.

Bonferroni correction. The family-wise significance threshold is $\alpha_{\text{Bonf}} = 0.05 \,/\, \binom{N}{2}$, applied uniformly across all pairwise comparisons. Pairs with $p < \alpha_{\text{Bonf}}$ are flagged as statistically significant.

Effect size. Cohen's d is computed for each ordered pair (m_1, m_2):

$$d = \frac{\bar{X}_{m_1} - \bar{X}_{m_2}}{s_p}, \qquad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}, \tag{3}$$

where s_i^2 is the sample variance of model i's score vector (ddof=1). The reverse pair (m_2, m_1) is assigned $d' = -d$, so the full $N \times N$ effect-size matrix is anti-symmetric.

Per-error-type statistics. The above pairwise statistics are also computed *independently for each taxonomy code* $e \in \{\text{E1}, \ldots, \text{E7}\}$, using the cluster-level binary column (1 if a row was assigned to that cluster, 0 otherwise) as the score variable. This yields a structured view of which model pairs differ significantly on specific error types rather than only globally, directly supporting the per-category analysis in Appendix I.

D.4. Output Schema

`agent_scores.csv`. Written per model folder. Contains all columns from the original `agent_response.csv` plus the four binary scores from Stage 1, the HDBSCAN `cluster` integer (-1 for noise), and the `error_type` string from Stage 2.

`model_evaluations.csv`. One row per `(model, metric, error_type)` triple. The `to` column is populated for pairwise rows and is `None` for per-model summary rows. Valid values of `metric` are: `Score`, `Refusal Score`, `Off Topic Score`, `Spatial Elements Score`, `Spatial Relations Score`, `Spatial Reasoning Score`, `bootstrap_mean`, `cohens_d`, `bonferroni_alpha`, `mcnemar_stat`.

model_evaluations.json. Nested dictionary keyed by model name. Top-level keys per model: `bootstrap_ci` (dict with `mean`, `ci_lower`, `ci_upper`, `std`), `average_score`, `overall_score`, `refusal_score`, `off_topic_score`, `spatial_elements_score`, `spatial_relations_score`, `spatial_reasoning_score`, `Cohen_d`, `Bonferroni`, `Mcnemar`. The last three keys are themselves dicts keyed by error type, each containing a list of `{"to": model_name, ...}` records.

D.5. Limitations and Design Trade-offs

LLM judge variance. Stage 1 scoring accuracy depends on the judge model's calibration on spatial reasoning tasks. We fix temperature at 0.1 and enforce strict JSON output to minimise inter-run variance, but absolute scoring accuracy has not been validated against human raters on this benchmark; we treat judge-assigned labels as high-precision proxies subject to the same caveats as any automated evaluator (**?**).

Clustering on small failure sets. HDBSCAN with minimum cluster size $k \geq 15$ requires at least ≈ 50 failed rows to produce at least two valid clusters. For models with very high accuracy (few failures), Stage 2 may return no clusters, leaving all rows with `error_type = None` and excluding them from per-type pairwise statistics. This is logged but not treated as an error.

Row alignment for pairwise statistics. McNemar's test and Cohen's d require instance-level alignment by `id`. If two models were not evaluated on an identical task set, pairwise statistics cover only the intersection; the intersection size is not currently surfaced in the output schema and should be verified before interpreting small-n pairings.

Known implementation bug. In `structured_signal_extraction`, `error_logs` is a `list` but is passed directly to `f.write()`, which raises a `TypeError` at runtime if any errors are logged. The fix is to replace `f.write(error_logs)` with `f.write("\n".join(error_logs))`. All results reported in this paper were produced on runs where no errors were logged, so this bug did not affect any reported scores.

E. Automated Error Analysis Pipeline

Benchmark papers in visual reasoning almost universally report accuracy summarised at the level of categories or aggregate datasets. Such tables are observational: they record *what* a model gets wrong but offer no mechanism for explaining *why*. This conflation of diverse failure mechanisms under a single accuracy figure obscures model behavior in ways that matter for both scientific understanding and practical model development.

FRONTIERSPATIAL's Automated Error Analysis Pipeline remedies this by supplying three scientific properties that do not currently coexist in any published spatial benchmark paper. *First*, a bottom-up inductive error taxonomy is derived from actual model failures rather than imposed from prior theoretical frameworks, making the taxonomy specific to the failure distributions observed in practice. *Second*, model comparison is conducted at the level of individual skill dimensions rather than aggregate accuracy, supported by appropriate inferential statistics. *Third*, the entire analysis harness is implemented as a fully reproducible, LLM-in-the-loop pipeline: every diagnostic claim produced by the system is directly traceable to structured data, and all intermediate artefacts are released alongside the final report so that every numerical finding is independently verifiable.

The pipeline consists of three sequential stages—STAGE 1: structured signal extraction; STAGE 2: error taxonomy induction; and STAGE 3: cross-model diagnostic synthesis—described in detail in the following subsections.

E.1. Stage 1: Structured Signal Extraction

Input format. STAGE 1 accepts as input either a JSON or CSV file with, at minimum, the fields `id`, `question`, `choices` (for multiple-choice questions), `answer` (ground truth), `prediction` (model output), `category`, and `sub_category`. An optional `reasoning_trace` field is consumed when chain-of-thought prompting was used during inference.

Scientific motivation. The standard error analysis in benchmark papers reduces every prediction to a binary *correct* / *incorrect* flag. This is observational, not diagnostic. Before failures can be classified by cause, it is necessary to extract structured signals from raw prediction strings that capture how the failure manifested, not merely that it occurred. Signal

extraction transforms each unstructured prediction record into a typed, multi-attribute object amenable to systematic downstream analysis.

E.1.1. SIGNAL S1: CORRECTNESS

For multiple-choice items, correctness reduces to a deterministic string match of the predicted choice letter against the gold answer. For open-ended (OE) items, correctness requires a semantic equivalence check, implemented via an LLM judge with a constrained prompt:

> *"Does the following prediction express the same spatial relationship as the gold answer? Answer* `yes`*/*`no` *and provide a one-sentence justification."*

The judge is invoked only for OE items scored as incorrect by string match—approximately 15–20% of benchmark items under the MCQ/OE balance from FRONTIERSPATIAL curation pipeline—making the approach cost-efficient without sacrificing accuracy on the cases where it matters.

E.1.2. SIGNAL S2: ANSWER FORMAT VALIDITY

The prediction string is parsed for structural validity independent of semantic correctness. Four output classes are distinguished: **valid_choice** (a recognisable answer token was produced), **refused** (the model explicitly declined to answer), and **off_topic** (the model produced a multi-sentence explanation in place of an answer). This classification is implemented purely with rule-based string analysis and requires no judge calls.

The diagnostic motivation for this separation is important: a refusal is a qualitatively different failure from a wrong answer. Conflating them in a single accuracy number obscures the distinction between models that attempt a spatial inference and fail versus models that decline to commit to an answer. The former reveals limits of spatial competence; the latter may reflect alignment training, calibration policies, or prompt sensitivity—none of which is informative about spatial reasoning capacity per se.

E.1.3. SIGNAL S3: REASONING TRACE COHERENCE

When a model was prompted with chain-of-thought, the extracted reasoning text is scored by a text-only LLM judge on three independent sub-criteria, each scored on the binary scale $\{0, 1\}$:

(a) **Element identification:** does the reasoning correctly identify the relevant spatial elements named in the question?

(b) **Relationship description:** is the spatial transformation or relationship correctly described in the reasoning steps?

(c) **Logical consistency:** does the stated conclusion follow logically from the stated reasoning?

The composite score $s_3 = \sum_{i=a}^{c} s_i \in \{0, 1, 2, 3\}$ supports detection of two diagnostically important failure modes that aggregate accuracy cannot distinguish. A model may arrive at a correct answer through internally inconsistent reasoning ($s_1 = 1$ but $s_3 < 3$), which constitutes a latent reasoning failure that will surface as errors under slight question reformulation. Conversely, a model may reason correctly but produce a malformed answer token ($s_1 = 0$ but $s_3 = 3$), which is a formatting failure rather than a spatial reasoning failure and should not penalise the model's competence estimate.

E.1.4. PRE-CLASSIFICATION INTO ERROR POOLS

The output of STAGE 1 is an enriched prediction file in which every item carries its full signal record. Incorrect items are automatically sorted into three mutually exclusive pools based on signal combinations, which direct STAGE 2's analysis budget toward the most informative failures:

- **Pool A** — *Format errors*: items where S2 produced a non-`valid_choice` label. These are format failures independent of spatial reasoning content.
- **Pool B** — *Reasoning errors*: items with a valid format (S2 passed) where the reasoning coherence score is $s_3 < 3$ (S3 failed). These are process failures where the model's answer-generation mechanism broke down despite producing a parseable output.

- **Pool C** — *Knowledge/perception errors*: items where S1 failed but S2 and S3 both passed. These are the most diagnostically rich failures: the model produced a valid, coherently reasoned answer that was nevertheless wrong.

Pool membership determines how STAGE 2's embedding and clustering are applied, ensuring that qualitatively different failure modes are not merged in a common semantic space during taxonomy induction.

E.2. Stage 2: Error Taxonomy Induction

Scientific motivation. Manually designed error taxonomies impose a theoretical structure that may or may not correspond to the actual distribution of failures produced by a given model on a given benchmark. Distinct models evaluated on the same benchmark can exhibit qualitatively different failure patterns, and a fixed *a priori* taxonomy risks either over-aggregating fine-grained patterns or missing failure modes that were not anticipated during taxonomy design. STAGE 2 adopts an inductive approach: the taxonomy is derived from the failures themselves, not from theory. This methodology is inspired by grounded theory coding in qualitative research and translated into an automated pipeline that combines embedding-based clustering with LLM-based labelling.

E.2.1. STEP 2A: ERROR EMBEDDING

For each incorrect item, a composite text string is constructed by concatenating the question, the gold answer, and the model prediction. This string is encoded using a strong text embedding model—either `e5-large-v2` or OpenAI `text-embedding-3-large`.

The key design decision is in what is embedded. The concatenation encodes the *failure pattern*—the semantic relationship between what was asked, what was correct, and what was predicted—rather than the question topic in isolation. As a consequence, clusters in the resulting embedding space reflect similar shapes of mistake rather than similar question domains. Two items from unrelated benchmark categories that elicit the same type of spatial confusion will be proximate in this space; two items from the same category that reflect different failure mechanisms will be distant. This property is what allows STAGE 2 to discover cross-cutting failure types that category-level accuracy tables cannot reveal.

Embeddings are computed separately for each error pool (**Pool A**, **Pool B**, **Pool C** from STAGE 1), so that format errors, reasoning errors, and knowledge errors cluster within their own semantic spaces rather than interfering with one another.

E.2.2. STEP 2B: DENSITY-BASED CLUSTERING

HDBSCAN (**?**) is applied to the per-pool embeddings in preference to K-means for two reasons grounded in the statistical properties of failure distributions. First, failure patterns are heterogeneous in density: some failure modes are highly consistent (tight geometric clusters reflecting a stereotyped error) while others are diffuse (scattered wrong answers with no single dominant pattern). HDBSCAN accommodates variable-density clusters without requiring a fixed global linkage threshold. Second, HDBSCAN explicitly designates noise points—items that do not belong to any cluster—corresponding to idiosyncratic failures that represent genuinely unusual model behavior rather than systematic error patterns.

The minimum cluster size hyperparameter is selected by a grid search over the range $\{10, 15, 20, 25, 30, 40, 50\}$, choosing the value that maximises the mean silhouette coefficient:

$$\bar{s} = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \frac{b(x) - a(x)}{\max(a(x), b(x))}, \tag{4}$$

where $a(x)$ is the mean intra-cluster distance for point x and $b(x)$ is the mean distance to the nearest cluster not containing x.

E.2.3. STEP 2C: LLM TAXONOMY LABELLING

For each discovered cluster, 10 representative items nearest to the cluster centroid are sampled and passed to an LLM with the following structured prompt:

> *"Here are 10 examples of a specific pattern of model failures on a spatial reasoning benchmark. Each shows the question, the correct answer, and the model's wrong prediction. Identify the common failure mode, give it a*

precise descriptive name (5 words maximum), and provide a one-sentence definition. The name should describe what the model did wrong, not the topic of the question."

This produces a human-readable label for each cluster that is grounded in actual failure instances rather than constructed from summary statistics. The constraint to name the failure mode rather than the question topic is critical: it prevents the LLM from producing category-level labels (e.g., "rotation questions") that would simply reproduce the existing benchmark taxonomy.

E.2.4. STEP 2D: MAPPING TO THE REFERENCE ERROR TAXONOMY

The inductively discovered clusters are mapped to a reference error taxonomy of seven types fixed *before* any model evaluation is conducted. This two-level approach—inductive discovery followed by mapping to a fixed schema—balances flexibility (the inductive step captures model-specific failure patterns) with comparability (the fixed schema enables consistent cross-model comparison). The seven reference types are:

E1 — Perceptual failure
The model misidentified or failed to detect a spatial element present in the image. The error occurs before any spatial inference is attempted.

E2 — Grounding failure
The model identified relevant visual elements but failed to anchor them in the correct spatial frame of reference (e.g., confused viewer-centric and object-centric reference frames).

E3 — Spatial transformation failure
The model understood the input spatial arrangement but failed to correctly apply a required rotation, reflection, or perspective change.

E4 — Quantitative spatial failure
Failure on tasks requiring counting, distance estimation, or spatial measurement, where the error is in quantification rather than spatial concept.

E5 — Multi-step reasoning failure
The model solved one spatial inference step correctly but failed on a subsequent step that required the output of the first step as input—a compositional failure rather than a single-step one.

E6 — Language–spatial mapping failure
The model misunderstood the spatial language in the question (e.g., confused "left of X" with "right of X", or misinterpreted a relative direction as absolute), producing an inference that is internally valid but addresses the wrong spatial relationship.

E7 — Format / refusal failure
The model did not produce a valid answer token regardless of spatial competence. This captures **Pool A** items that reach the taxonomy stage.

The LLM mapping from discovered clusters to reference types is validated by a human reviewer on a 50-item stratified sample, computing inter-rater agreement via Cohen's κ. A target of $\kappa \geq 0.75$ is required before the labeled dataset is accepted for use in STAGE 3; items on which the LLM mapping is rejected by the reviewer are relabelled manually and logged in the validation report.

Scientific contribution of Stage 2. The central finding that STAGE 2 enables—and that no per-category accuracy table can reveal—is that error types do not partition cleanly along benchmark category lines. A model may fail on a category such as *3D Pose* for reasons distributed across **E1**, **E3**, and **E5** simultaneously, while failing on *Camera Motion* predominantly through **E2**. This cross-cutting structure is invisible in standard benchmark tables and constitutes a substantive empirical finding about the relationship between spatial task structure and model failure mode.

E.3. Stage 3: Cross-Model Diagnostic Synthesis

Scientific motivation. A single model's error profile is descriptive. Error profiles compared consistently across a population of 20–25 models are *generative*—they identify which error types are universal (every model fails on **E3** at comparable rates, indicating genuine benchmark difficulty) versus model-specific (one model fails disproportionately on **E2** while another fails on **E5**, revealing architectural or training-data differences). STAGE 3 is the analytical engine that transforms N individual error profiles into cross-cutting findings about the current state of spatial reasoning in vision-language models.

E.3.1. STEP 3A: PER-MODEL SKILL PROFILING

Each model's error labels from STAGE 2 are aggregated into a 7-dimensional *error profile vector* $\mathbf{p} \in [0,1]^7$, where p_k is the proportion of that model's failures attributable to error type **Ek**. Complementarily, a *skill accuracy vector* $\mathbf{a} \in [0,1]^7$ is computed, where a_k is the model's accuracy on all items for which **Ek** is the primary challenge. These two views are informationally distinct: the error distribution describes what the model gets wrong; the skill accuracy describes what it can do on the hardest instances of each sub-competency.

Both vectors are computed at four levels of granularity: (i) overall, (ii) per main benchmark category, (iii) per sub-category, and (iv) per difficulty bin using the difficulty ratings from FRONTIERSPATIAL curation pipeline. All outputs are stored as structured JSON per model to support programmatic downstream analysis and automated report generation.

E.3.2. STEP 3B: STATISTICAL RIGOR LAYER

Three statistical procedures are applied systematically to ensure that comparative claims have quantifiable inferential support:

McNemar's test for paired item comparison. For each pair of models (M_i, M_j), McNemar's test is applied to the binary correctness vectors across all shared items. This tests whether the two models make *statistically different errors on the same items*, which is more informative than comparing aggregate accuracy: two models may have identical overall accuracy yet systematically fail on different items, indicating complementary rather than equivalent spatial competencies.

Bootstrap confidence intervals. Confidence intervals on each model's per-error-type accuracy are computed via nonparametric bootstrap with $B = 1{,}000$ resampling iterations. This produces exact 95% confidence intervals without distributional assumptions, which is appropriate given the heterogeneous sample sizes across error types and benchmark categories.

Effect size with multiple-comparison correction. For each pair of models on each skill dimension, Cohen's d is computed as the effect size. Significance is reported at $\alpha = 0.05$ with Bonferroni correction for the number of simultaneous pairwise comparisons. This correction ensures that the reported p-values account for the family-wise error rate under the full comparison matrix. A representative example of a reportable finding takes the form:

> *"Model A significantly outperforms Model B on **E3** (spatial transformation), Cohen's $d = 0.43$, $p < 0.001$ (Bonferroni-corrected)."*

This formulation grounds every diagnostic claim in a specific statistical result that is fully reproducible from the released data.

E.3.3. STEP 3C: LLM DIAGNOSTIC SYNTHESIS

To produce a human-readable, actionable report rather than a data dump, each model's full profile is passed to an LLM under a structured prompt that includes: the per-category accuracy table, the 7-dimensional error distribution, the top-3 most frequent cluster labels from STAGE 2, cross-model rank on each skill dimension, and the statistical contrast results from Step 3b. The LLM synthesises this into 3–5 structured findings, each expressed in the canonical form:

> **Finding:** [claim in plain language]. **Evidence:** [specific numbers and error types]. **Implication:** [what this reveals about the model's spatial reasoning capacity].

Each finding is at most 2–3 sentences in the claim component, but every numerical assertion in the evidence component is drawn exclusively from the structured data passed to the LLM—it cannot hallucinate statistics it was not given. The

entire chain from raw prediction file to narrative finding is logged, and all intermediate files are released, so every claim is independently verifiable by inspection.

E.4. Output Artefacts

STAGE 3 produces three complementary output types, together constituting a complete, citable diagnostic report.

Per-model skill card. For each evaluated model, a skill card is generated in both JSON and Markdown formats comprising: (i) accuracy across **E1–E7** summarised as a 7-dimensional profile suitable for radar chart visualisation; (ii) per-category accuracy ranked from strongest to weakest; (iii) the model's top-5 failure clusters from STAGE 2 with representative example items; and (iv) a *blind-spot flag* for any error type on which the model scores more than one standard deviation below the cross-model population mean—a simple signal to model developers and downstream users of where the model's spatial reasoning is structurally deficient.

Cross-model comparison table. A summary table with rows indexing models and columns indexing skill dimensions. Each cell contains the model's accuracy on that dimension, its rank position among all evaluated models, and a significance marker indicating whether the difference from the best-performing model on that dimension is statistically significant at the Bonferroni-corrected $\alpha = 0.05$ level.

Narrative synthesis report. A field-level report produced by the LLM synthesis step in Step 3c, summarising: which error types are universally hard across all models (high mean failure rate, low variance); which skill dimensions show the most inter-model variance and therefore discriminate best between models; and which models exhibit unusual profiles suggesting distinctive training or architectural properties. This report is designed to be incorporated directly into benchmark papers as a section on cross-model analysis.

E.5. Reproducibility and Infrastructure Considerations

The pipeline is designed with reproducibility as a first-order constraint, not a post-hoc property. All intermediate artefacts produced by STAGE 1–STAGE 3—the enriched signal files, the cluster assignments and centroids, the LLM judge call logs with full prompts and responses, the statistical test outputs, and the narrative synthesis inputs—are released alongside the benchmark dataset and model outputs. This means that every numerical claim in a paper relying on the pipeline is independently verifiable by re-running the relevant downstream analysis step from the released files.

Detailed code and running details are provided with the submission.

Model-agnostic incremental updates. The pipeline is intentionally structured so that its most expensive component—STAGE 2 taxonomy induction, which involves embedding all failure items and running HDBSCAN—needs to be run only once for an initial model cohort. When a new model is added to the evaluation, its failure items are embedded and assigned to the existing cluster structure using nearest-centroid assignment; the taxonomy induction step is not re-run. The statistical comparison in STAGE 3 automatically updates to include the new model. This makes the harness a *living* evaluation infrastructure: it remains useful and consistent as new models are released, exactly the persistent evaluation architecture required for a benchmark intended to track progress in spatial reasoning over time.

Cost profile. The LLM calls in STAGE 2 (taxonomy labelling) and STAGE 3 (diagnostic synthesis) are bulk operations over cluster representatives rather than per-item calls, and their cost is therefore sub-linear in the number of items. For a benchmark of $N = 5{,}000$ items evaluated across 25 models, the total LLM inference cost of the full pipeline is estimated to be around $400-500 at current API rates, making it accessible for research groups without large compute budgets.

F. More details on Experiments and Evaluation

We use `google/gemini-3.1-flash-image-preview` the judge model (including error taxonomy construction). We set `temperature` to 0.1. For the taxonomy generation stage, we use a maximum of 250 tokens, while for both the answering model and judge model we use a maximum of 1000 tokens. All other generation parameters are kept at their default values.

Prompts are available in Appendix **??**.

F.1. Evaluation Protocol

Judging Protocol. All responses are evaluated using a separate vision-language judge model under a strict binary scoring scheme. The judge compares the model output against the ground-truth answer, allowing only trivial surface variations (e.g., whitespace, capitalization), while rejecting semantically incomplete or approximate responses. Correct answers embedded within longer generations are accepted, whereas any ambiguity is conservatively scored as incorrect. The judge returns a structured JSON object containing the binary score and a short justification.

Decoding and Inference Settings. We set the generation temperature to 0.1 to minimize stochasticity and ensure stable outputs across model variants. Maximum generation length is set to 600 tokens for standard models and 20,000 tokens for reasoning-augmented models, allowing sufficient capacity for multi-step reasoning when required. The judge model is allocated a maximum of 1,200 tokens. All evaluations are performed via the OPENROUTER API using default settings otherwise. API usage costs range from approximately \$0.10 to \$15 per 1M input/output tokens, with total experimental expenditure (including pilot runs and ablations) of approximately \$1,200.

Rationale. This evaluation design isolates semantic correctness from superficial variations in formatting and verbosity, ensuring that measured differences reflect reasoning ability rather than presentation style. The strict binary judge enforces consistency across models, enabling stable and reproducible comparisons under invariant output transformations.

F.2. Evaluation of LLM-as-a-Judge for Open-ended Evaluation

To evaluate *Gemini-3.1-Flash-Image-Preview* as an automatic judge, we construct a balanced evaluation set of 175 samples spanning all task categories. Each sample consists of a question, ground-truth label, and model-generated response, and is independently annotated by three human evaluators using binary judgments (1 for correct, 0 for incorrect).

We compare the LLM judgments against human annotations using multiple agreement metrics computed both globally and per category. Specifically, we report Cohen's κ, accuracy against the majority vote, and inter-annotator agreement measured via average pairwise κ among human annotators.

Cohen's Kappa. Cohen's κ quantifies agreement between two raters while correcting for chance-level alignment. It is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o denotes observed agreement and p_e denotes expected agreement under independence:

$$p_e = \sum_{c \in \{0,1\}} p_c^{(a)} \, p_c^{(b)}.$$

Here, $p_c^{(a)}$ and $p_c^{(b)}$ are the empirical probabilities that rater a and rater b assign label c. In our setup, we compute κ between the LLM judge and the majority-vote label derived from human annotations, denoted as $\kappa_{\text{judge-human}}$.

Accuracy with Majority Vote. We also report direct agreement with the majority label. Let y_i^{LLM} denote the LLM prediction and y_i^{maj} the majority vote. Accuracy is computed as

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(y_i^{\text{LLM}} = y_i^{\text{maj}}\right).$$

Unlike Cohen's κ, this metric does not adjust for chance agreement and measures only exact label correspondence.

Inter-Annotator Agreement. To quantify consistency among human annotators, we compute the average pairwise Cohen's κ across all annotator pairs. Let $\mathcal{H} = \{H_1, H_2, H_3\}$ denote the set of annotators. The inter-annotator agreement is defined as

$$\kappa_{\text{human}} = \frac{1}{|\mathcal{P}|} \sum_{(a,b) \in \mathcal{P}} \kappa(a, b),$$

where $\mathcal{P}$ is the set of all unordered annotator pairs. This provides a robust estimate of annotation consistency without assuming a fixed reference label.

Higher values indicate stronger agreement beyond chance, while lower values reflect increased ambiguity or subjectivity in the task.

Table 10. Agreement metrics between the LLM judge and human annotators across categories. The table reports accuracy and Cohen's κ between the judge and human majority vote, along with average pairwise human κ as a measure of annotator consistency.

Category	n	Accuracy	$\kappa_{\text{judge-human}}$	$\kappa_{\text{human-pairwise}}$
Overall	175	**1.000**	**1.000**	0.805
Physical Reasoning	25	1.000	1.000	0.739
Positional & Geometric Reasoning	25	1.000	1.000	0.827
Knowledge-Grounded Reasoning	25	1.000	1.000	0.853
Relational - Logical & Causal Reasoning	25	1.000	1.000	0.819
Temporal & Dynamic Reasoning	25	1.000	1.000	0.719
Quantitative & Comparative Reasoning	25	1.000	1.000	0.781
Perceptual Understanding	25	1.000	1.000	**0.922**

Interpretation. Table 10 shows that the LLM judge achieves perfect agreement with human majority annotations across all task types, with both accuracy and Cohen's κ equal to 1.000. This indicates that the judge fully replicates the aggregated human decision boundary without any observed disagreement across the 175 evaluated samples.

However, a more nuanced picture emerges when examining inter-annotator agreement. Pairwise human κ varies substantially across task types, ranging from 0.719 (Temporal & Dynamic Reasoning) to 0.922 (Perceptual Understanding). This spread highlights that while the judge is perfectly aligned with the majority vote, the underlying human annotations themselves are not uniformly consistent. Tasks involving temporal dynamics and physical reasoning exhibit comparatively lower agreement, suggesting higher ambiguity or interpretive variance among annotators.

In contrast, categories such as Perceptual Understanding ($\kappa = 0.922$) and Knowledge-Grounded Reasoning ($\kappa = 0.853$) demonstrate strong human consistency, indicating clearer decision boundaries. The fact that the LLM judge achieves perfect agreement even in lower-consensus categories suggests that it effectively captures the majority aggregation rule rather than independently resolving ambiguity.

These findings indicate that the LLM-as-judge framework is highly reliable as a proxy for majority human evaluation, achieving deterministic alignment across all categories. However, perfect κ should be interpreted cautiously: it reflects agreement with the majority vote rather than independent judgment quality. The variation in human pairwise agreement underscores that certain reasoning categories remain intrinsically subjective, and future evaluation frameworks may benefit from explicitly modeling annotator uncertainty rather than collapsing it into a single ground truth.

Overall, the results validate the use of LLM-based automated judges for scalable evaluation, while also emphasizing the importance of complementary human agreement metrics to contextualize performance.

F.3. Human Evaluation Protocol

To establish a human performance baseline, we evaluate the same set of questions using three independent participants: one high-school-level annotator, one STEM undergraduate, and one non-STEM undergraduate (at most in their second year of study). Each participant independently answers all questions without access to external tools or prior solutions.

Responses are scored using the same binary evaluation protocol as model outputs using LLM as a Judge, where answers are marked as correct (1) or incorrect (0) based on exact semantic equivalence with the ground truth, allowing only minor formatting variations. No collaboration or discussion is permitted during evaluation.

Final human labels are obtained via majority vote across the three annotators. Majority voting is used to reduce individual-level variance and annotation noise, yielding a more stable estimate of typical human performance. While this does not

eliminate systematic biases shared across annotators, it provides a robust aggregate baseline for comparison against model performance.

F.4. AI Agents for Spatial Reasoning

F.4.1. SPATIOXOLVER

We include SPATIOXOLVER (Wasi et al., 2026), a multi-agent spatial reasoning system adapted from Xolver (Hosain et al., 2025), to evaluate the effectiveness of structured decomposition on challenging instances. The system factorizes visual reasoning into a sequence of specialized agents, each responsible for a distinct stage of the perception–reasoning pipeline.

The pipeline begins with a *visual analysis agent*, which produces a detailed scene description capturing objects, attributes, and coarse spatial layout. An *object segmentation agent* then assigns consistent identifiers to individual entities, followed by an *attribute extraction agent* that encodes properties (e.g., shape, size, color, orientation) in structured form. A *spatial relation agent* constructs pairwise and higher-order relations (e.g., left-of, above, alignment), forming a relational graph of the scene. Building on this, a *grouping and symmetry agent* identifies higher-level structures such as clusters, grids, and symmetries, while a *transformation tracking agent* captures changes across frames for dynamic reasoning tasks.

All intermediate outputs are consolidated into a unified structured representation, which is consumed by a final *reasoning agent* to generate answers. This modular design enables explicit multi-step reasoning by separating perception, grounding, and compositional inference, and allows the system to propagate low-level visual signals into higher-order spatial reasoning. In our evaluation, SPATIOXOLVER serves as a representative agentic approach for assessing whether such decomposition can recover performance on instances unsolved by monolithic VLMs.

F.4.2. CLAUDE CODE AGENT

We also evaluate a **Claude Code-based agent** (Anthropic, 2026), which performs spatial reasoning through sequential programmatic decomposition. Instead of distributing reasoning across specialized agents, this approach leverages iterative code generation and execution to structure intermediate reasoning steps. The agent dynamically constructs representations (e.g., lists, counters, relational checks) within a single reasoning loop, enabling symbolic manipulation and step-wise verification. This design is particularly effective for tasks requiring arithmetic reasoning, rule-based logic, or structured comparison, where explicit computation can reduce ambiguity.

F.4.3. EVALUATION SETUP

We evaluate two agentic systems on a *hard-case subset* consisting of 175 instances that were not correctly solved by any of the 16 evaluated frontier VLMs. Both systems receive identical inputs (image and question) and are evaluated under the same protocol as standard models.

SPATIOXOLVER is implemented using *Gemini-3.1-Flash-Preview* as the underlying model, with a low temperature (0.1) to ensure deterministic behavior. It employs a multi-agent pipeline that decomposes each task into structured sub-components (e.g., object extraction, attribute encoding, relation construction, and transformation reasoning), producing a unified intermediate representation before generating the final answer. In contrast, the **Claude Code-based agent** performs sequential reasoning within a single loop, leveraging programmatic execution to construct and verify intermediate representations.

For both systems, we allow extended reasoning with a maximum context length of 10,000 tokens to support multi-step decomposition and intermediate computation. Each agent produces a single final answer per instance, which is evaluated using the same LLM-based judging framework described in Section 3.2. No external retrieval or additional supervision is provided beyond the input image and question. This setup ensures a controlled comparison, isolating the impact of multi-agent decomposition versus sequential programmatic reasoning on solving previously unsolved spatial reasoning tasks.

G. Model Performance Breakdown on FRONTIERSPATIAL

G.1. Model Performance On Identifying Spatial Relations

Table 11 shows a clear pattern is the uneven distribution of strengths across categories, with most models achieving higher scores in Quantitative and Knowledge reasoning while lagging in Physical and Temporal reasoning, indicating that symbolic

or count-based tasks remain easier than dynamic or embodied spatial understanding. Among proprietary systems, GEMINI-3.1-FLASH-LITE leads Physical reasoning (65.68%), whereas CLAUDE HAIKU 4.5 dominates Positional/Geometric (60.40%), Temporal (78.90%), and Quantitative (86.28%), suggesting a more balanced but still category-skewed capability profile. In contrast, GPT-4O-MINI consistently underperforms across all categories, rarely exceeding 55%, highlighting a substantial gap between smaller optimized variants and stronger mid-tier models like GPT-5-MINI and O4-MINI.

Open-source models show greater variance but also stronger peaks, with QWEN3.5-9B and QWEN3.6-35B-A3B matching or exceeding proprietary leaders in multiple categories, particularly Knowledge (75%+) and Quantitative (≈75%), indicating competitive reasoning depth at higher parameter scales. Notably, KIMI K2.5 achieves the highest Relational/Logical score (69.98%) and strong Temporal (72.02%) and Quantitative (80.72%) performance, reflecting robustness in multi-step and dependency-heavy reasoning tasks. A consistent trend across both groups is the relative difficulty of Temporal reasoning, where even top models show larger variance (roughly 48% to 79%), implying instability in handling dynamic spatial changes.

Perceptual understanding exhibits a bifurcation: weaker models cluster around 50–60%, while stronger open models like QWEN3.5-9B (78.67%) and KIMI K2.5 (77.49%) achieve substantially higher accuracy, suggesting that visual grounding scales effectively with model capacity. Overall, the results reveal that no single model dominates all categories, and performance leadership shifts depending on the reasoning type, reinforcing the claim that spatial reasoning is multi-faceted and cannot be captured by aggregate scores alone.

Table 11. Model Performance (%) on Identifying Spatial Relations across Categories.

Model	**Phys.** (#1312)	**Pos.Geo.** (#2596)	**Knowledge** (#2332)	**Re.Lo.Ca.** (#455)	**Temporal** (#218)	**Quant.** (#503)	**Percept.** (#422)
Proprietary Models							
Claude Haiku 4.5	50.99	**60.40**	71.40	64.84	**78.90**	**86.28**	69.19
Gemini-2.5-Flash	44.66	53.21	62.72	58.02	48.62	55.98	56.64
Gemini-3.1-Flash-Lite	**65.68**	59.83	67.11	59.12	62.39	69.58	65.17
GPT-4o-mini	34.53	40.57	50.21	51.43	41.28	53.48	50.24
GPT-5-mini	49.70	49.15	65.34	60.22	56.88	61.63	56.64
o4-mini	49.39	55.10	**58.99**	**60.66**	61.93	67.79	**56.87**
Open-Source Models							
Gemma-3-4B-it	37.27	45.21	41.68	44.62	21.56	62.62	56.64
Qwen3-VL-8B-Instruct	42.99	51.04	55.00	53.41	56.88	63.42	52.37
Qwen3.5-9B	**65.68**	55.66	75.09	64.84	66.97	**75.75**	**78.67**
Llama-3.2-11B-Vision-Ins.	33.92	39.74	45.20	46.59	36.70	48.71	46.21
Gemma-4-31B-it	57.55	54.75	69.00	59.12	61.93	65.21	61.37
Aya-Vision-32B	40.63	41.31	33.50	44.27	56.88	63.62	54.74
Qwen3.6-35B-A3B	62.22	**57.56**	**75.56**	66.74	**68.40**	75.55	71.29
Kimi K2.5	55.18	58.29	73.36	**69.98**	72.02	80.72	77.49
GLM-4.5V	41.39	56.87	62.93	62.86	61.93	60.64	60.81
Mistral Small 2603	38.72	40.69	55.36	48.57	55.50	57.46	46.45

G.2. Model Performance On Identifying Spatial Element

The overall pattern in Table 12 shows uniformly high performance across models, with most scores concentrated between 75% and 90%, indicating that spatial element identification is substantially easier than spatial relation reasoning and exhibits lower variance across categories. Proprietary models are tightly clustered at the top, with CLAUDE HAIKU 4.5 achieving the highest Physical (93.67%) and Temporal (91.28%) scores, while GEMINI-3.1-FLASH-LITE leads Positional/Geometric (73.81%) and Relational/Logical (89.89%), reflecting strong consistency across structurally diverse tasks. In contrast to Table 1, weaker models such as GPT-4O-MINI still maintain relatively high performance (≈45–80%), suggesting that element identification relies less on complex reasoning chains and more on direct visual-text alignment.

Open-source models display both competitiveness and slight dispersion, with QWEN3.5-9B, QWEN3.6-35B-A3B, and KIMI K2.5 achieving top-tier results across multiple categories, particularly in Perceptual (up to 89.10%) and Temporal (up

to 83.49%), in some cases exceeding proprietary counterparts. A notable trend is the compression of category gaps: unlike spatial relation tasks, Temporal and Physical categories no longer exhibit extreme drops, indicating that identifying elements in dynamic or physical contexts is less sensitive to reasoning difficulty. Quantitative performance is consistently high across nearly all models (≈78–88%), reinforcing that counting or magnitude extraction remains a stable capability.

Perceptual understanding shows the highest ceiling overall, with several models surpassing 85%, suggesting that visual feature extraction and localization scale effectively with both model size and multimodal training. However, Positional/Geometric reasoning still exhibits moderate spread (≈54–78%), implying that even element identification retains some sensitivity to spatial structure complexity. Overall, the results indicate that while model differences persist, spatial element identification is largely a saturated capability, and the primary bottleneck for MLLMs lies in transforming these identified elements into coherent relational reasoning.

Table 12. Model Performance (%) on Identifying Spatial Elements across Categories.

Model	**Phys.** (#1312)	**Pos.Geo.** (#2596)	**Knowledge** (#2332)	**Re.Lo.Ca.** (#455)	**Temporal** (#218)	**Quant.** (#503)	**Percept.** (#422)
Proprietary Models							
Claude Haiku 4.5	**93.67**	72.81	87.48	89.67	**91.28**	88.47	87.44
Gemini-2.5-Flash	85.14	63.02	81.98	83.52	54.13	75.70	75.83
Gemini-3.1-Flash-Lite	88.49	**73.81**	87.78	**89.89**	80.28	82.50	84.36
GPT-4o-mini	79.95	56.12	81.39	77.80	45.41	76.14	77.49
GPT-5-mini	86.89	62.43	**88.50**	84.18	65.14	**88.47**	78.20
o4-mini	87.88	64.20	79.37	85.49	68.81	86.08	**79.86**
Open-Source Models							
Gemma-3-4B-it	85.90	57.50	72.43	73.41	39.45	83.50	81.99
Qwen3-VL-8B-Instruct	77.82	65.22	81.90	80.00	65.60	79.72	79.38
Qwen3.5-9B	88.86	76.35	84.48	86.15	80.73	80.12	**89.10**
Llama-3.2-11B-Vision-Ins.	85.29	56.91	76.16	79.56	46.79	86.08	71.80
Gemma-4-31B-it	**89.18**	69.52	86.62	**89.01**	81.65	78.13	80.57
Aya-Vision-32B	85.60	64.01	66.54	76.65	63.30	88.47	75.83
Qwen3.6-35B-A3B	88.51	**78.02**	86.23	87.00	80.19	85.69	85.41
Kimi K2.5	88.41	72.65	83.66	87.64	**83.49**	**84.69**	**87.91**
GLM-4.5V	84.60	72.47	81.77	83.96	77.06	84.10	81.47
Mistral Small 2603	83.61	53.96	**82.38**	77.14	58.72	79.52	70.62

G.3. Model Performance On Spatial Reasoning

Performance on spatial relation identification in Table 11 is both lower and more dispersed than element identification, with scores ranging roughly from 20% to 80%, indicating higher task complexity and greater sensitivity to reasoning capability. A consistent pattern is that Quantitative and Knowledge categories yield comparatively higher scores across models, while Physical and Temporal reasoning exhibit larger drops and variance, reflecting difficulty in handling dynamic or grounded spatial interactions. Among proprietary systems, CLAUDE HAIKU 4.5 leads Positional/Geometric (60.40%), Temporal (78.90%), and Quantitative (86.28%), whereas GEMINI-3.1-FLASH-LITE achieves the top Physical score (65.68%), highlighting complementary strengths rather than a single dominant model.

Lower-tier proprietary models such as GPT-4O-MINI show consistently reduced performance across all categories, often below 55%, reinforcing the dependence of relational reasoning on model capacity and reasoning depth. Open-source models exhibit a wider spread but also stronger peaks, with QWEN3.5-9B and QWEN3.6-35B-A3B achieving top Knowledge (≈75%) and Quantitative (≈75%) scores, matching or surpassing proprietary counterparts. KIMI K2.5 stands out in Relational/Logical reasoning (69.98%) and maintains strong Temporal (72.02%) and Perceptual (77.49%) performance, indicating robustness in multi-step relational inference.

Temporal reasoning remains the most unstable category, with scores spanning from 21.56% (GEMMA-3-4B-IT) to 78.90% (CLAUDE HAIKU 4.5), suggesting inconsistent handling of time-dependent spatial relations. Perceptual performance shows a clear scaling trend, where stronger open models such as QWEN3.5-9B achieve high accuracy (78.67%), while smaller

models cluster near 50–60%. Overall, no model consistently dominates across all categories, and the shifting leadership across tasks underscores that spatial relation reasoning is heterogeneous and not reducible to a single capability dimension.

Table 13. Model Performance (%) on Spatial Reasoning across Categories.

Model	**Phys.** (#1312)	**Pos.Geo.** (#2596)	**Knowledge** (#2332)	**Re.Lo.Ca.** (#455)	**Temporal** (#218)	**Quant.** (#503)	**Percept.** (#422)
Proprietary Models							
Claude Haiku 4.5	37.88	32.66	48.93	48.13	**61.47**	**18.69**	45.97
Gemini-2.5-Flash	37.88	35.96	42.81	49.01	39.91	11.95	45.50
Gemini-3.1-Flash-Lite	**50.23**	**43.48**	51.37	50.11	57.34	16.50	**51.42**
GPT-4o-mini	29.27	24.10	23.03	42.42	21.10	13.52	40.28
GPT-5-mini	43.67	36.51	**52.60**	55.38	55.50	16.50	48.10
o4-mini	42.99	36.39	47.45	**56.70**	54.13	16.50	46.21
Open-Source Models							
Gemma-3-4B-it	27.21	25.13	13.16	31.87	11.01	8.95	38.39
Qwen3-VL-8B-Instruct	35.67	36.08	31.45	44.84	48.17	15.51	41.23
Qwen3.5-9B	30.13	27.20	31.86	36.92	38.07	11.33	41.47
Llama-3.2-11B-Vision-Ins.	25.30	24.34	17.32	36.48	20.18	9.15	33.65
Gemma-4-31B-it	**49.09**	35.17	31.73	45.93	54.13	15.51	50.47
Aya-Vision-32B	25.93	22.16	15.57	35.02	29.36	15.31	41.71
Qwen3.6-35B-A3B	40.48	35.89	**45.98**	48.02	38.21	13.92	43.06
Kimi K2.5	35.52	34.45	39.90	42.60	44.50	12.13	44.79
GLM-4.5V	33.46	**39.46**	37.41	**50.33**	**50.46**	**14.91**	**49.41**
Mistral Small 2603	29.65	25.21	29.33	36.04	49.08	14.51	32.94

H. Category-wise Model Performance of FRONTIERSPATIAL

H.1. Closed-source Models

Sub-Category	Task	Claude Haiku 4.5	Gemini 2.5 Flash	Gemini 2.0 Flash	Mistral Small	o4-mini	GPT-4o mini	GPT-5 mini
Abstract Knowledge	Riddle	**5.48**	1.37	2.74	2.74	2.74	0.00	4.11
	Pattern	18.64	20.34	16.95	20.34	15.25	**25.42**	15.25
	Trick	78.00	76.00	**84.00**	62.00	76.00	66.00	80.00
Domain Specific	Scientific Reasoning	64.00	56.00	68.00	22.00	**72.00**	26.00	68.00
	Engineering & Technology	29.41	13.73	33.33	35.29	**41.18**	33.33	37.25
Mathematical Reasoning	Angle Reasoning	63.64	54.23	**68.53**	38.46	55.94	30.07	60.56
	Circle Geometry	59.09	47.73	**62.12**	25.76	53.79	18.94	61.36
	Shape Decomposition	55.61	42.93	**61.95**	30.73	47.80	24.39	53.17
	Metric Geometry	59.93	47.43	**61.40**	38.24	53.14	25.00	**61.40**
	Geometric Construction	44.21	29.47	47.37	30.53	45.26	21.05	**50.53**
	Math – Arithmetic	60.75	46.73	58.88	27.10	**61.68**	19.63	60.75
	Geometry – Descriptive	30.61	27.55	30.61	25.51	34.69	19.39	**40.82**
	Geometry – Coordinate	53.25	46.75	55.84	27.27	59.74	36.36	**68.83**
	Scaling & Ratio	56.64	55.94	**65.73**	34.97	55.94	25.87	52.45
	Math – Algebra	54.85	42.81	51.51	21.40	60.87	17.39	**64.21**
	Geometry – Transformational	31.85	33.76	**49.68**	29.30	32.48	19.75	40.13
	Constraint Geometry	57.47	45.98	**68.97**	33.33	48.28	26.44	52.87
	Geometric Reasoning	33.33	29.49	34.62	26.92	30.77	20.51	**41.03**

Table 14. Open-ended Evaluation Accuracy on Knowledge-Grounded Reasoning Category by Question Sub-Categories for Closed-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Claude Haiku 4.5	Gemini 2.5 Flash	Gemini 2.0 Flash	Mistral Small	o4-mini	GPT-4o mini	GPT-5 mini
Object-Level Perception	Size & Scale	45.88	47.65	**55.88**	40.00	46.47	43.53	55.29
	Attribute	**18.87**	5.66	13.21	11.32	7.55	13.21	3.77
	Exist or Object Identification	52.00	74.00	**76.00**	40.00	72.00	60.00	66.00
	Color	56.00	56.00	64.00	48.00	**66.00**	44.00	64.00
	Appearance	2.04	**8.16**	4.08	**8.16**	6.12	6.12	2.04
Scene-Level Perception	Scene Understanding	**76.00**	74.00	72.00	32.00	74.00	**76.00**	72.00

Table 15. Open-ended Evaluation Accuracy on Perceptual Understanding Category by Question Sub-Categories for Closed-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Claude Haiku 4.5	Gemini 2.5 Flash	Gemini 2.0 Flash	Mistral Small	o4-mini	GPT-4o mini	GPT-5 mini
Navigation & Control	Turning – Car	31.71	37.80	**50.00**	24.39	26.83	14.63	48.78
	Navigation	48.02	54.80	55.37	37.29	51.41	43.50	**55.93**
	Lane Position	37.28	43.42	58.33	30.70	**64.91**	32.46	55.26
	Turning – Entity	29.44	27.10	**35.98**	29.91	26.64	25.23	31.31
	Lane Change	23.89	22.78	**56.11**	18.89	22.22	18.33	22.78
Physical Interaction	Property Inference	51.35	**75.68**	74.32	32.43	67.57	47.30	70.27
	Collision	10.47	14.53	18.02	10.47	**20.93**	7.56	16.28
Predictive Physics	Predictive	40.00	38.00	**44.00**	42.00	36.00	32.00	38.00
	Planning	25.93	33.33	**38.52**	18.52	22.22	23.70	28.15

Table 16. Open-ended Evaluation Accuracy on Physical Reasoning Category by Question Sub-Categories for Closed-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Claude Haiku 4.5	Gemini 2.5 Flash	Gemini 2.0 Flash	Mistral Small	o4-mini	GPT-4o mini	GPT-5 mini
Counting	Counting	35.53	32.00	42.11	22.37	36.84	31.58	**50.00**
	Relational Counting	36.00	38.00	32.00	**68.00**	32.00	46.00	22.00
Measurement	Relative Distance	**8.55**	2.56	7.69	2.56	3.85	2.56	4.70
	Measurement	**36.07**	13.11	18.03	24.59	**36.07**	22.95	29.51
	Absolute Distance	0.00	1.22	1.22	0.00	**7.32**	1.22	3.66

Table 17. Open-ended Evaluation Accuracy on Quantitative & Comparative Reasoning Category by Question Sub-Categories for Closed-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Claude Haiku 4.5	Gemini 2.5 Flash	Gemini 2.0 Flash	Mistral Small	o4-mini	GPT-4o mini	GPT-5 mini
Causal Reasoning	Logical Counterfactual	46.02	32.74	41.59	19.47	**57.52**	24.78	**57.52**
	Physical Causality	53.23	**59.68**	40.32	**62.90**	53.23	53.23	50.00
Relations	Spatial Relationship	47.50	53.93	55.36	37.50	55.71	47.86	**56.07**

Table 18. Open-ended Evaluation Accuracy on Relational, Logical & Causal Reasoning Category by Question Sub-Categories for Closed-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Claude Haiku 4.5	Gemini 2.5 Flash	Gemini 2.0 Flash	Mistral Small	o4-mini	GPT-4o mini	GPT-5 mini
2D Reasoning	2D Geometry	29.41	31.37	**43.14**	32.35	40.20	24.51	34.31
3D Reasoning	3D Pose	20.35	26.11	**32.30**	23.45	30.09	22.57	29.20
	3D Folding	36.76	**79.41**	77.94	19.12	29.41	42.65	55.88
	3D Object Detection	3.31	**20.53**	6.62	1.32	7.95	3.97	1.99
	3D Rotation	27.92	27.92	**34.52**	27.92	32.49	24.37	30.46
	3D Assembly	9.24	**10.92**	9.24	8.40	8.40	**10.92**	9.24
	3D Projection	**32.00**	**32.00**	**32.00**	20.00	16.00	24.00	26.00
	3D Shape Finding	21.43	**44.64**	41.07	25.00	28.57	17.86	25.00
	3D Geometry	48.85	42.41	**58.45**	38.11	49.00	31.81	52.15
Geometric Reasoning	Direction and Orientation	51.03	57.93	54.48	51.03	57.24	45.52	**60.00**
	Occluder Attributes	15.47	32.04	**48.07**	14.36	28.18	13.26	32.60
	Compatibility	51.14	55.68	48.86	39.77	**59.09**	50.00	54.55
	Occluded Attributes	18.98	35.56	**53.21**	19.25	31.82	11.50	36.36
Localization	Object Localization	76.00	**90.00**	86.00	58.00	86.00	80.00	82.00
	Relative Location	66.00	78.00	**90.00**	26.00	70.00	54.00	70.00
Positional Relations	(Region – Region)	35.44	24.05	29.11	15.19	**36.71**	18.99	16.46
	(Camera – Region)	**35.80**	22.22	24.69	8.64	28.40	18.52	19.75
	(Camera – Object)	23.26	22.09	19.77	19.77	**33.72**	20.93	20.93
	(Object – Object)	21.84	10.34	16.09	16.09	19.54	17.24	**20.69**
	(Object – Region)	**29.82**	7.02	14.04	8.77	24.56	7.02	–

Table 19. Open-ended Evaluation Accuracy on Spatial & Geometric Reasoning Category by Question Sub-Categories for Closed-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Claude Haiku 4.5	Gemini 2.5 Flash	Gemini 2.0 Flash	Mistral Small	o4-mini	GPT-4o mini	GPT-5 mini
Motion Reasoning	Camera Motion	14.29	15.71	20.00	32.86	**40.00**	31.43	28.57
	Object Motion	**80.00**	56.00	76.00	70.00	70.00	22.00	**80.00**
Temporal Reasoning	Temporal Movement	**37.25**	27.45	27.45	21.57	25.49	17.65	21.57
	Sequencing	**100.00**	80.85	95.74	89.36	95.74	19.15	**100.00**

Table 20. Open-ended Evaluation Accuracy on Temporal & Dynamic Reasoning Category by Question Sub-Categories for Closed-source Models. Best score per row is bold and underlined.

H.2. Open-source Models

Sub-Category	Task	Gemma 3 4B	Gemma 4 31B	Llama 3.2 11B	Qwen3 VL 8B	Qwen3.5 9B
Creative Abstract Knowledge	Riddle	1.37	1.37	2.74	1.37	1.37
	Pattern	8.47	20.34	18.64	13.56	8.47
	Trick	58.00	80.00	50.00	68.00	66.00
Domain Specific	Scientific Reasoning	14.00	46.00	24.00	40.00	50.00
	Engineering & Technology	27.45	29.41	23.53	23.53	21.57
Mathematical Reasoning	Angle Reasoning	11.89	20.28	18.88	35.21	25.17
	Circle Geometry	9.09	19.70	15.91	34.09	18.18
	Shape Decomposition	8.78	18.05	17.56	38.54	15.61
	Metric Geometry	5.51	16.18	21.32	34.56	19.49
	Geometric Construction & Optimization	15.79	20.00	21.05	28.42	16.84
	Math – Arithmetic	15.89	**53.27**	11.21	33.64	43.93
	Geometry – Descriptive	10.20	28.57	11.22	26.53	17.35
	Geometry – Coordinate	15.58	25.97	16.88	36.36	38.96
	Scaling & Ratio	9.09	14.69	20.98	32.17	15.38
	Math – Algebra	4.68	32.78	10.70	31.44	32.11
	Geometry – Transformational	19.11	24.84	19.75	27.39	10.83
	Constraint Geometry	10.34	22.99	20.69	35.63	18.39
	Geometric Reasoning & Visual Problem Solving	13.25	28.63	16.67	24.79	11.97

Sub-Category	Task	Qwen3.6 35B	Aya Vision 32B	Kimi K2.5	GLM 4.5V
Creative &Abstract Knowledge	Riddle	4.11	1.37	**4.11**	2.74
	Pattern	11.86	20.34	10.17	**27.12**
	Trick	74.00	66.00	**78.00**	69.39
Domain Specific	Scientific Reasoning	56.00	16.00	**62.00**	54.00
	Engineering & Technology	25.49	23.53	29.41	**33.33**
Mathematical Reasoning	Angle Reasoning	**46.76**	16.08	38.73	31.47
	Circle Geometry	**37.69**	14.39	34.85	32.58
	Shape Decomposition	**39.90**	14.63	33.17	25.37
	Metric Geometry	**44.94**	16.91	34.93	33.09
	Geometric Construction & Optimization	**40.00**	16.84	24.21	30.53
	Math – Arithmetic	50.47	16.82	50.47	**53.27**
	Geometry – Descriptive	**29.59**	18.37	14.29	28.57
	Geometry – Coordinate	**52.00**	16.88	42.86	40.26
	Scaling & Ratio	**44.37**	12.59	40.56	32.87
	Math – Algebra	45.82	11.74	**50.84**	45.82
	Geometry – Transformational	**37.58**	21.02	18.47	27.39
	Constraint Geometry	28.74	19.54	32.18	**40.23**
	Geometric Reasoning & Visual Problem Solving	27.78	17.95	21.79	**26.07**

Table 21. Open-ended Evaluation Accuracy on Knowledge-Grounded Reasoning Category by Question Sub-Categories for Open-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Gemma 3 4B	Gemma 4 31B	Llama 3.2 11B	Qwen3 VL 8B	Qwen3.5 9B
Object-Level Perception	Size & Scale	39.41	**55.88**	35.88	40.59	35.29
	Attribute	**15.09**	**15.09**	3.77	**15.09**	9.43
	Exist or Object Identification	38.00	**76.00**	46.00	60.00	54.00
	Color	36.00	52.00	58.00	50.00	46.00
	Appearance	**10.20**	2.04	8.16	**10.20**	4.08
Scene-Level Perception	Scene Understanding	72.00	**82.00**	54.00	68.00	76.00

Sub-Category	Task	Qwen3.6 35B	Aya Vision 32B	Kimi K2.5	GLM 4.5V
Object-Level Perception	Size & Scale	40.72	47.65	43.53	**54.12**
	Attribute	13.21	11.32	9.43	7.55
	Exist or Object Identification	70.00	36.00	**76.00**	69.39
	Color	51.02	46.00	50.00	**66.00**
	Appearance	4.08	**10.20**	8.16	6.12
Scene-Level Perception	Scene Understanding	66.00	76.00	64.00	70.00

Table 22. Open-ended Evaluation Accuracy on Perceptual Understanding Category by Question Sub-Categories for Open-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Gemma 3 4B	Gemma 4 31B	Llama 3.2 11B	Qwen3 VL 8B	Qwen3.5 9B
Navigation & Control	Turning – Car	23.17	**54.88**	36.59	53.66	6.10
	Navigation	35.03	51.98	37.85	40.11	35.03
	Lane Position	26.32	67.54	19.74	39.91	65.79
	Turning – Entity	25.23	**38.32**	21.03	34.58	2.35
	Lane Change	18.89	**42.78**	12.22	38.89	0.56
Physical Interaction	Property Inference	32.43	66.22	40.54	62.16	56.76
	Collision	11.05	13.95	11.05	2.33	3.49
Predictive Physics	Predictive	36.00	40.00	36.00	28.00	16.00
	Planning	17.78	31.11	26.67	28.89	20.00

Sub-Category	Task	Qwen3.6 35B	Aya Vision 32B	Kimi K2.5	GLM 4.5V
Navigation & Control	Turning – Car	20.73	20.73	9.76	31.71
	Navigation	42.37	**44.07**	**43.50**	**43.50**
	Lane Position	**72.81**	22.54	51.75	39.47
	Turning – Entity	22.50	18.69	18.22	22.43
	Lane Change	15.00	7.78	40.56	28.33
Physical Interaction	Property Inference	**67.57**	52.70	64.86	62.16
	Collision	11.63	**19.19**	12.21	9.30
Predictive Physics	Predictive	28.00	30.00	**40.00**	34.00
	Planning	32.84	20.77	22.22	**25.93**

Table 23. Open-ended Evaluation Accuracy on Physical Reasoning Category by Question Sub-Categories for Open-source Models (Part 1: Gemma 3 4B – Qwen3.5 9B; Part 2: Qwen3.6 35B – GLM 4.5V). Best score per row is bold and underlined.

Sub-Category	Task	Gemma 3 4B	Gemma 4 31B	Llama 3.2 11B	Qwen3 VL 8B	Qwen3.5 9B
Counting	Counting	21.05	**40.79**	11.84	28.95	28.95
	Relational Counting	6.00	48.00	28.00	64.00	26.00
Measurement	Relative Distance	1.28	4.70	1.71	3.85	0.00
	Measurement	29.51	18.03	27.87	18.03	4.92
	Absolute Distance	1.22	0.00	0.00	2.44	1.22

Sub-Category	Task	Qwen3.6 35B	Aya Vision 32B	Kimi K2.5	GLM 4.5V
Counting	Counting	32.89	18.42	28.95	26.32
	Relational Counting	24.00	**70.00**	12.00	36.00
Measurement	Relative Distance	3.42	2.99	**6.41**	0.43
	Measurement	18.03	34.43	11.48	**47.54**
	Absolute Distance	1.22	0.00	3.66	**4.88**

Table 24. Open-ended Evaluation Accuracy on Quantitative & Comparative Reasoning Category by Question Sub-Categories for Open-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Gemma 3 4B	Gemma 4 31B	Llama 3.2 11B	Qwen3 VL 8B	Qwen3.5 9B
Causal Reasoning	Logical Counterfactual	12.39	22.12	13.27	30.97	18.58
	Physical Causality	58.06	41.94	**62.90**	46.77	37.10
Relations	Spatial Relationship	35.00	52.86	41.07	49.64	36.07

Sub-Category	Task	Qwen3.6 35B	Aya Vision 32B	Kimi K2.5	GLM 4.5V
Causal Reasoning	Logical Counterfactual	30.97	12.39	**38.94**	34.51
	Physical Causality	48.39	**72.58**	30.00	**62.90**
Relations	Spatial Relationship	51.25	36.56	41.07	**54.64**

Table 25. Open-ended Evaluation Accuracy on Relational, Logical & Causal Reasoning Category by Question Sub-Categories for Open-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Gemma 3 4B	Gemma 4 31B	Llama 3.2 11B	Qwen3 VL 8B	Qwen3.5 9B
2D Reasoning	2D Geometry	23.53	**33.33**	24.51	23.53	9.80
3D Reasoning	3D Pose	17.70	29.65	17.26	**30.09**	20.35
	3D Folding	16.18	45.59	23.53	52.94	8.82
	3D Object Detection	12.58	17.88	19.21	**22.52**	5.96
	3D Rotation	30.46	25.89	**38.58**	31.98	11.17
	3D Assembly	5.04	7.56	**12.61**	10.92	0.00
	3D Projection	20.00	**32.00**	20.00	20.00	2.00
	3D Shape Finding	30.36	26.79	26.79	23.21	3.57
	3D Geometry	22.06	41.79	22.92	35.53	26.65
Geometric Reasoning	Direction and Orientation	38.62	**57.93**	42.76	41.38	43.45
	Occluder Attributes	19.89	30.94	20.44	44.20	39.23
	Compatibility	42.05	**62.50**	43.18	51.14	52.27
	Occluded Attributes	18.45	29.14	20.32	43.85	39.04
Localization	Object Localization	64.00	90.00	74.00	86.00	80.00
	Relative Location	44.00	68.00	60.00	76.00	60.00
Positional Relations	Pos. Rel. (Region – Region)	**29.11**	27.85	21.52	21.52	2.53
	Pos. Rel. (Camera – Region)	22.22	16.05	16.05	22.22	7.41
	Pos. Rel. (Camera – Object)	20.93	23.26	9.30	19.77	10.47
	Pos. Rel. (Object – Object)	11.49	24.14	9.20	18.39	3.45
	Pos. Rel. (Object – Region)	21.05	21.05	7.02	10.53	5.26

Sub-Category	Task	Qwen3.6 35B	Aya Vision 32B	Kimi K2.5	GLM 4.5V
2D Reasoning	2D Geometry	14.71	24.51	8.82	29.41
3D Reasoning	3D Pose	25.66	21.68	**30.09**	27.88
	3D Folding	42.65	26.47	14.71	**55.88**
	3D Object Detection	18.54	7.95	0.66	10.60
	3D Rotation	24.46	27.92	5.58	30.96
	3D Assembly	9.32	10.92	2.54	10.08
	3D Projection	14.00	18.00	6.00	30.00
	3D Shape Finding	30.36	17.86	3.57	**39.29**
	3D Geometry	41.79	25.57	35.63	**42.69**
Geometric Reasoning	Direction and Orientation	52.41	41.38	50.00	48.97
	Occluder Attributes	49.17	9.39	**61.33**	39.23
	Compatibility	52.27	55.68	33.33	57.95
	Occluded Attributes	48.40	9.89	**63.64**	42.25
Localization	Object Localization	80.00	68.00	82.00	**92.00**
	Relative Location	**78.00**	44.00	70.00	68.00
Positional Relations	Pos. Rel. (Region – Region)	1.27	25.32	7.69	25.32
	Pos. Rel. (Camera – Region)	8.64	24.69	20.99	**37.04**
	Pos. Rel. (Camera – Object)	16.28	11.63	17.44	**40.70**
	Pos. Rel. (Object – Object)	8.05	10.34	17.24	**28.74**
	Pos. Rel. (Object – Region)	3.51	19.30	15.79	**31.58**

Table 26. Open-ended Evaluation Accuracy on Positional & Geometric Reasoning Category by Question Sub-Categories for Open-source Models. Best score per row is bold and underlined.

Sub-Category	Task	Gemma 3 4B	Gemma 4 31B	Llama 3.2 11B	Qwen3 VL 8B	Qwen3.5 9B
Motion Reasoning	Camera Motion	4.29	10.00	**18.57**	**18.57**	2.86
	Object Motion	16.00	70.00	24.00	**72.00**	60.00
Temporal Reasoning	Temporal Movement	21.57	**33.33**	23.53	19.61	0.00
	Sequencing	0.00	**97.87**	29.79	85.11	95.74

Sub-Category	Task	Qwen3.6 35B	Aya Vision 32B	Kimi K2.5	GLM 4.5V
Motion Reasoning	Camera Motion	14.29	4.29	7.14	11.43
	Object Motion	42.00	44.00	**70.00**	**70.00**
Temporal Reasoning	Temporal Movement	26.67	25.49	3.92	21.57
	Sequencing	48.94	42.55	80.85	91.49

Table 27. Open-ended Evaluation Accuracy on Temporal & Dynamic Reasoning Category by Question Sub-Categories for Open-source Models. Best score per row is bold and underlined.

I. Qualitative Error Taxonomy of Spatial–Visual Reasoning Failures

To systematically characterise the failure modes observed across all evaluated models, we constructed a two-level hierarchical error taxonomy. The *first level* (E1–E7) partitions failures by the cognitive or computational operation that breaks down. The *second level* decomposes each first-level category into focused subcategories that distinguish the representational domain, modality, or reasoning substrate in which the failure manifests.

Taxonomy construction proceeded inductively: every failed instance was first assigned a first-level code based on *where in the perception–reasoning pipeline* the model deviated from the correct answer; it was then assigned a second-level code based on the *domain or object type* involved. Where a single instance exhibited multiple failure signatures (e.g., a driving task requiring both perceptual identification and egocentric grounding), it was assigned to the category corresponding to the earliest breakdown step. The resulting taxonomy covers 7 first-level categories and 38 second-level subcategories, spanning 13 evaluated models.

Table 28 provides a compact summary; the remainder of this section defines each category in detail.

Table 28. Overview of the two-level spatial–visual error taxonomy.

Code	Name	Brief Definition
E1	Perceptual Failure	Misidentification or non-detection of a visual element
E2	Spatial Grounding Failure	Elements identified but not anchored to the correct spatial frame
E3	Spatial Transformation Failure	Arrangement understood but rotation/folding/perspective incorrectly applied
E4	Quantitative Spatial Failure	Counting, distance estimation, or measurement errors
E5	Multi-Step Reasoning Failure	Correct first step but failure on a dependent subsequent step
E6	Language–Spatial Mapping Failure	Misinterpretation of spatial language (*e.g.*, left/right confusion)
E7	Format / Refusal Failure	No valid answer produced, regardless of spatial competence

I.1. E1 — Perceptual Failure

Definition. A **perceptual failure** occurs when the model misidentifies, misclassifies, or entirely fails to detect a visual or spatial element that is unambiguously present in the image. The failure precedes any spatial reasoning: the model's internal scene representation is incorrect from the outset, so downstream spatial inferences are compromised regardless of their logical validity.

Benchmark examples.

- *"What is the colour of the water in the image?"* — model reports the wrong colour despite a clearly visible body of water.
- *"What is the shape of the brick in the picture when viewed from above?"* — model cannot identify the top-view silhouette of a LEGO brick.
- *"What kind of animal is on top of the tree?"* — model hallucinate an absent animal or misidentifies the species present.
- *"What is the primary type of scene depicted in this aerial image?"* — model confuses an aerial residential scene with an industrial or natural landscape.
- *"Which chest X-ray disease can be found?"* — model fails to identify a pathological feature that a radiologist would detect.

E1.1 — Object & Attribute Recognition Failure.
Failure to correctly identify an object's perceptual attributes such as colour, shape, or semantic category.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` ,`GLM-4.5V`.

Root-cause analysis. Object attribute recognition errors are among the most broadly distributed failures in the taxonomy, appearing across nearly every evaluated model. Smaller-parameter models (`Qwen3-VL-8B-Instruct`

, Llama-3.2-11B-Vision-Instruct) lack sufficient visual encoder capacity to disentangle fine-grained attribute channels—colour, texture, and shape—from the holistic patch embeddings produced by their vision backbones. Mid-size models (Mistral-Small-2603, Gemma-4-31B-it) show a pattern consistent with inadequate cross-modal alignment training: the language head defaults to high-frequency category priors (e.g., assigning the most common LEGO brick colour in the training corpus) rather than grounding in the specific visual instance. Instruction-tuned models (GPT-4o-mini, o4-mini) occasionally hallucinate plausible-but-incorrect attributes, consistent with the known tendency of RLHF-trained models to generate fluent, overconfident descriptions. Even Gemini-2.5-Flash exhibits residual confusion on low-contrast or partially occluded objects, pointing to limits in pre-training image diversity rather than model capacity per se.

E1.2 — Scene & Environment Classification Failure.
Failure to correctly classify the scene type, environment category, or contextual setting (e.g., aerial view, indoor layout, road type).

Affected models: Gemini-3.1-Flash-Preview, Gemma-4-31B-it, Mistral-Small-2603, Kimi K2.5, Qwen3-VL-8B-Instruct ,GLM-4.5V.

Root-cause analysis. Scene-level classification demands the integration of holistic contextual cues (skyline geometry, road markings, vegetation density) that are poorly preserved by the reduced-resolution encoders used in efficiency-oriented models such as Gemini-3.1-Flash-Preview and Qwen3-VL-8B-Instruct . Kimi K2.5 and GLM-4.5V are pre-trained predominantly on East-Asian data distributions; aerial and architectural scenes sourced from other geographic contexts may lie outside their training support, causing systematic misclassification of unfamiliar regional layouts. Gemma-4-31B-it and Mistral-Small-2603 appear to rely heavily on salient foreground objects for scene inference, neglecting background context entirely, which leads to errors whenever the dominant foreground category is ambiguous with respect to scene type. None of the affected models appears to employ an explicit multi-scale scene integration mechanism, leaving them vulnerable to tasks that require simultaneous reasoning about local objects and global environmental structure.

E1.3 — Driving-Scene Perceptual Failure.
Failure to correctly perceive lane positions, vehicle turning behaviour, safety- relevant distances, or other road-level visual cues.

Affected models: Gemma-3-4B-it, Gemma-4-31B-it, Gemini-3.1-Flash-Preview, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, o4-mini, GPT-5-mini, Qwen3-VL-8B-Instruct , GLM-4.5V.

Root-cause analysis. Driving-scene perception requires simultaneous tracking of road markings, vehicle silhouettes, and implicit motion cues within a single static frame—a high-density, context-sensitive task that challenges all evaluated models. The smallest models (Gemma-3-4B-it, Qwen3-VL-8B-Instruct) cannot reliably resolve lane boundary markings from road-surface texture at standard benchmark image resolutions due to limited encoder feature density. GLM-4.5V and Gemini-3.1-Flash-Preview frequently conflate structurally similar lane configurations (e.g., merge zones versus standard two-lane roads), suggesting over-reliance on template matching rather than fine-grained geometric parsing of road topology. Llama-3.2-11B-Vision-Instruct and Mistral-Small-2603 exhibit elevated false-negative rates on turning-vehicle detection, likely because dynamic turning-manoeuvre imagery is underrepresented in their visual instruction-tuning corpora relative to static or parked-vehicle scenarios. Even o4-mini shows residual errors on ambiguous turning-angle estimation, highlighting a fundamental gap between static-image perception and the temporal motion understanding available in video-based training paradigms.

E1.4 — Action & Social Outcome Perception Failure.
Failure to correctly detect or infer an agent's action, intent, or predicted social or physical outcome from visual evidence.

Affected models: Gemini-3.1-Flash-Preview, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, o4-mini, GPT-4o-mini, GPT-5-mini.

Root-cause analysis. Predicting human actions from a static frame is intrinsically under-determined, requiring strong priors about body pose, gaze direction, and contextual social norms that are difficult to learn from standard image–caption pre-training. Llama-3.2-11B-Vision-Instruct and Gemini-3.1-Flash-Preview default to the statistically most common action label (e.g., "standing" rather than "reaching") when perceptual evidence is ambiguous, revealing frequency-

based prior dominance over evidence-driven inference. GPT-4o-mini and GPT-5-mini occasionally construct internally coherent but visually unsupported action narratives—a hallmark of language-prior dominance in RLHF-trained models where fluency incentives outweigh perceptual fidelity. Mistral-Small-2603's failures cluster on social-outcome forecasting tasks (e.g., predicting the result of a physical interaction), suggesting that its instruction-tuning data contained limited diversity in social-scene annotations. o4-mini's errors are concentrated on fine-grained limb-movement detection, pointing to the upper resolution limit of its vision encoder for body-part localisation under natural viewing conditions.

E1.5 — Visual Pattern & Embedded Figure Detection Failure.
Failure to detect hidden figures, visual regularities, or spatially encoded cues embedded within complex or cluttered images.

Affected models: Gemini-3.1-Flash-Preview, Llama-3.2-11B-Vision-Instruct, GPT-4o-mini, GPT-5-mini.

Root-cause analysis. Embedded-figure detection requires figure-ground segregation and contour grouping across distractors—a Gestalt-level perceptual operation that is poorly supported by patch-based vision encoders optimised for natural-image classification. Llama-3.2-11B-Vision-Instruct and Gemini-3.1-Flash-Preview show near-chance performance on shape-embedding tasks, indicating that their encoders do not produce sufficiently separated figure–ground representations to isolate target shapes from complex backgrounds. GPT-4o-mini and GPT-5-mini perform better but fail specifically when the embedded figure shares colour or texture statistics with its surround, suggesting that their contour-grouping abilities remain partially dependent on luminance-contrast cues rather than purely topological shape properties. None of the four models appears to implement an iterative visual search strategy; each commits to a single feedforward perceptual interpretation, which is systematically insufficient for heavily camouflaged or overlapping figure tasks.

E1.6 — Symbolic & Language-Embedded Visual Failure.
Failure on tasks requiring interpretation of formally rendered symbols, written puzzles, logic diagrams, or code-like visual structures embedded in images.

Affected models: Claude 4.5 Haiku, Gemini-2.5-Flash, Gemini-3.1-Flash-Preview, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, Kimi K2.5, o4-mini, GPT-4o-mini, GPT-5-mini, Qwen3-VL-8B-Instruct ,GLM-4.5V.

Root-cause analysis. Visually rendered formal notation (automata diagrams, logic gates, arithmetic puzzles) requires joint low-resolution OCR and domain-specific semantic rule application—a compound task for which general-purpose vision-language pre-training provides limited supervision. Across all affected models the most common error is OCR-level misreading of small, stylised, or densely packed characters (e.g., digit–letter confusion in state-machine labels), propagating into downstream logical failures. Lighter-weight models (Qwen3-VL-8B-Instruct ,Gemini-3.1-Flash-Preview) additionally lack exposure to specialised formal-notation corpora, resulting in near-random responses on finite-automaton and logic-gate tasks. The breadth of this failure even in large models (Gemini-2.5-Flash, o4-mini) confirms that visually rendered formal notation represents a persistent weak spot for current vision–language architectures, likely requiring targeted data curation rather than scale-alone remediation.

I.2. E2 — Spatial Grounding Failure

Definition. A **spatial grounding failure** occurs when the model correctly identifies the elements present in a scene but cannot anchor them to their correct positions within a spatial frame of reference. The model has an adequate symbolic inventory of the scene yet fails to assign the correct spatial roles (left/right, above/below, near/far, inside/outside) or cannot resolve which reference frame—egocentric vs. allocentric—the question presupposes.

Benchmark examples.

- *"How would you describe the lane position of Entity #1? front lane / front-left lane / front-right lane / oncoming traffic lane?"* —model names the correct vehicle but assigns it the wrong lane.
- *"Are the LEGO pieces pointed to by the two red arrows adjoining or separated?"* — model recognises both bricks but misjudges their contact relationship.

- *"If the fully visible 'one way' sign is facing south, what direction is the other 'one way' sign facing?"* — model identifies both signs but applies the wrong allocentric transformation.
- *"When a person exits the door with a semicircular glass top facing south, which room is on the west side of the corridor?"* — model navigates the verbal description but grounds the directions incorrectly.
- *"Where is the bridge located in the image?"* — model identifies the bridge but mis-reports its spatial relation to the scene.

E2.1 — Egocentric / First-Person Frame Grounding Failure.
Failure to ground objects or directions relative to the observer's or ego-vehicle's viewpoint.

Affected models: all 13 evaluated models.

Root-cause analysis. Egocentric frame grounding is the single most broadly distributed failure in the taxonomy, appearing in every evaluated model without exception. Vision–language models are pre-trained predominantly on image–caption pairs that use allocentric (world-centred) spatial language, leaving egocentric descriptions (*"to my left," "approaching me"*) statistically underrepresented in the training distribution. During instruction-tuning, egocentric driving questions are frequently resolved by heuristic pattern-matching to common caption templates rather than by explicit viewpoint transformation, producing systematic left–right and forward–backward reversals. The fact that even the largest evaluated models (`Gemini-2.5-Flash`, `GPT-5-mini`) exhibit this error confirms that parameter scale alone cannot compensate for the distributional gap in egocentric spatial supervision. Models not specifically fine-tuned on autonomous-driving datasets are particularly affected: without explicit training signal that distinguishes the ego-vehicle's first-person perspective from a third-person observer's view, the model cannot reliably switch reference frames on demand.

E2.2 — Relative Object Positioning Failure.
Failure to correctly specify where one identified object is positioned relative to another (left/right, above/below, in-front/behind).

Affected models: all 13 evaluated models.

Root-cause analysis. Relative positional grounding requires three sequential steps: (1) identify the reference object, (2) identify the target object, and (3) compute and verbalise the directional relationship between them—a chain that introduces compounding error at each stage. In smaller models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`), the dominant failure is reference-object selection: the model anchors on the most salient or centrally positioned object rather than the one designated by the question. Mid-size models (`Mistral-Small-2603`, `Kimi K2.5`) more often select the correct reference object but err during directional computation, particularly when the reference and target objects are near-equidistant from the image centre or when viewpoint changes are involved. The universality of this error across all parameter scales underscores that relative spatial reasoning is a core linguistic–perceptual interface problem requiring explicit compositional supervision that is absent from standard next-token or captioning objectives.

E2.3 — Depth & Layering Grounding Failure.
Failure to correctly assign depth order, occlusion relationships, or vertical-dimension stacking among objects.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Mistral-Small-2603`.

Root-cause analysis. Depth ordering from monocular images requires simultaneous integration of multiple monocular depth cues: occlusion boundaries, perspective foreshortening, texture gradients, and relative size—a multi-cue fusion problem for which single-frame vision encoders offer only weak inductive biases. `Gemini-3.1-Flash-Preview` and `Mistral-Small-2603` rely disproportionately on the size-as-depth heuristic, causing systematic errors whenever a physically closer object is smaller than a farther one. `Claude 4.5 Haiku`'s errors concentrate on vertical-stack tasks (e.g., layered cross-section diagrams), suggesting that its encoder conflates image-plane vertical position with world-space depth when objects are stacked top-to-bottom in the image rather than arranged in perspective. `Gemini-2.5-Flash`'s depth errors arise primarily on visibility-and- occlusion queries where partial occlusion is the sole available depth cue, pointing to an under-representation of partial-occlusion grounding annotations in its pre-training data.

E2.4 — Road & Lane Spatial Grounding Failure.

Failure to correctly anchor lanes, navigation paths, or individual traffic participants to their precise road-level positions.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` ,`GLM-4.5V`.

Root-cause analysis. Road-level spatial grounding demands both fine-grained lane-boundary perception and a compositional map of which entity occupies which lane—a joint perceptual and relational reasoning requirement. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) fail at the perceptual stage: lane markings are not reliably resolved at their native encoder resolutions, so the grounding problem is never properly posed. Mid-size models (`Mistral-Small-2603`, `GLM-4.5V`) resolve the markings but fail to consistently map entities to their correct lanes, especially under multi-entity scenes where multiple vehicles occupy adjacent lanes simultaneously. Larger models (`GPT-4o-mini`, `o4-mini`) show residual confusion specifically on oncoming-traffic lane assignment, likely because their training data contains relatively few right-hand-drive versus left-hand-drive scene variants, causing occasional frame-of-reference inversion.

E2.5 — LEGO & Structured-Assembly Grounding Failure.
Failure to correctly ground LEGO brick positions, adjacency relationships, or assembly placement within a 3D structural frame.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` ,`GLM-4.5V`.

Root-cause analysis. LEGO grounding tasks require precise stud-level spatial accounting in an isometric 3D render where perspective foreshortening compresses spatial relationships, making millimetre-scale adjacency judgements necessary. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) lack sufficient patch resolution to reliably distinguish stud-facing surface from body-plate surface in isometric renders. Mid-size models produce more accurate brick inventories but lose track of inter-brick contact surfaces when scenes contain more than three to four bricks, suggesting a limited spatial working-memory capacity within a single forward pass. Even the largest evaluated models (`GPT-5-mini`, `Gemini-2.5-Flash`) make grounding errors on adjacency tasks, confirming that fine-grained 3D contact geometry remains a challenging perceptual problem at all evaluated parameter scales in the absence of dedicated 3D-supervision pre-training.

E2.6 — Agent Action & Social Grounding Failure.
Failure to correctly ground an agent's action or predicted behaviour to the correct spatial participant or scene context.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Llama-3.2-11B-Vision-Instruct`, `GLM-4.5V`.

Root-cause analysis. Action grounding requires the model to bind a detected action to the correct spatial agent when multiple agents are present in the same scene—a cross-referential operation that combines visual entity tracking with predicate-argument structure. `Llama-3.2-11B-Vision-Instruct` and `Gemini-3.1-Flash-Preview` frequently misassign an observed action to the wrong agent, particularly in scenes where two agents are spatially proximate or visually similar. `Claude 4.5 Haiku` and `GLM-4.5V` show systematic confusion when the question uses deictic expressions (*"the man on the horse"*) that require resolving a chain of visual references before the target agent is unambiguously identified. The error is exacerbated by the absence of explicit agent-tracking supervision in standard image–caption pre-training, where captions describe scene contents globally rather than establishing agent-specific spatial bindings.

I.3. E3 — Spatial Transformation Failure

Definition. A **spatial transformation failure** occurs when the model has correctly perceived and grounded the spatial arrangement of elements but fails to correctly apply a mental operation—rotation, folding, reflection, perspective shift, or physical state change—to produce the answer. The failure is not in *what* the model sees, but in *what it computes* from that scene.

Benchmark examples.

- *"Which image in the following options shows the LEGO object in the reference image rotated clockwise by 120 degrees?"*

— model identifies the correct object but selects the wrong rotation option.

- *"Work out which option shows the figure on the left when folded along the dotted line."* — model perceives the 2D net correctly but produces an incorrect 3D fold result.
- *"If the billiard ball moves in the direction of the red arrow, will any ball go into the pocket and score?"* — model understands the arrow direction but incorrectly predicts the physical trajectory.
- *"Figure 1 shows the north side of the object. Which side does Figure 2 show?"* — model applies an incorrect compass-direction transformation.
- *"A person hangs a very heavy, swinging hammock from the edge of the overhanging roof. What is the most likely outcome?"* — model fails to apply physical causality reasoning to the described scenario.

E3.1 — 2D Folding & Net Transformation Failure.
Failure to mentally fold a 2D net into a 3D shape or predict the correct result of a paper-folding operation.

Affected models: `Claude 4.5 Haiku`, `Gemini-3.1-Flash-Preview`, `Gemma-4-31B-it`, `Mistral-Small-2603`, `o4-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` .

Root-cause analysis. Net-folding tasks require the simultaneous tracking of face adjacency, face orientation, and edge-pairing across a complete 2D-to-3D transformation—a demanding spatial-composition operation with no natural analogue in image captioning or language modelling. `Gemma-4-31B-it` and `Mistral-Small-2603` exhibit failures concentrated on octahedral and irregular-polygon nets, where the mapping from 2D face layout to 3D face identity is non-obvious and requires multi-step mental rotation before any face-matching can begin. `Qwen3-VL-8B-Instruct` and `Gemini-3.1-Flash-Preview` fail even on cubic nets, suggesting that their encoders do not produce the edge-connectivity representations needed to reason about fold topology. `o4-mini` and `Claude 4.5 Haiku` succeed on standard cube-folding variants but fail on higher-complexity nets (dodecahedra, irregular polyhedra), indicating that their folding competence is pattern-matched to common training instances rather than generalisable to novel net configurations. None of the affected models appears to decompose the folding task into sub-steps (identify faces → label edges → simulate fold), instead attempting a holistic one-shot transformation that is too complex for reliable execution.

E3.2 — 3D Rotation & Isometric View Failure.
Failure to mentally rotate a 3D object or correctly match it to a given isometric, top-down, or perspective view.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. 3D mental rotation is the most taxonomically widespread transformation failure, appearing in twelve of thirteen evaluated models and spanning the full range of parameter scales. Fine-angle discrimination (30° vs. 60° clockwise rotation of a LEGO model viewed from above) requires stable internal 3D representations that are not produced by patch-based encoders trained on 2D images. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) lack the feature resolution to track individual stud positions across rotation, making fine-angle discrimination near-impossible. Mid-size models (`Mistral-Small-2603`, `GLM-4.5V`) succeed at coarse 90° rotations but fail at 30° and 60°, suggesting that their mental rotation competence is driven by symmetry-based shortcuts rather than genuine continuous rotation. Even `GPT-5-mini` and `o4-mini` exhibit above-chance but below-human performance, confirming that current vision–language training objectives do not adequately incentivise the construction of view-invariant 3D object representations.

E3.3 — Physical Causality & Trajectory Prediction Failure.
Failure to predict the physical state or trajectory of an object after applying a force, motion, or physical interaction.

Affected models: `Claude 4.5 Haiku`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Llama-3.2-11B-Vision-Instruct`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`.

Root-cause analysis. Physical causality prediction requires an internal approximate physics engine that can simulate Newtonian dynamics from a single static image—a capability absent from standard discriminative or autoregressive training objectives. `Gemini-3.1-Flash-Preview` and `Gemma-3-4B-it` fail on simple billiard- ball trajectory tasks, indicating

that even two-body Newtonian interactions exceed their causal reasoning capacity when posed in a purely visual format. `Llama-3.2-11B-Vision-Instruct` produces physically plausible but locally incorrect trajectories (e.g., correct direction but wrong target pocket), suggesting a shallow heuristic that matches surface-level arrow cues without integrating collision geometry. Instruction-tuned models (`GPT-4o-mini`, `GPT-5-mini`) show language-prior dominance: they predict the "most common" physical outcome for the described scenario rather than reading the specific geometric configuration in the image, leading to failures on edge cases. `Claude 4.5 Haiku` and `o4-mini` succeed on standard collisions but fail on multi-body or structural-load scenarios (e.g., hammock attachment physics), implying that their causal competence does not generalise beyond the most common physical interaction patterns seen during training.

E3.4 — Driving & Ego-Motion Transformation Failure.
Failure to transform driving-scene observations into a correct prediction of ego-vehicle motion, lane-change outcome, or vehicle turning trajectory.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. Ego-motion prediction requires composing lane-position grounding with a forward simulation of vehicle kinematics—a two-stage transformation that compounds error from both perceptual and reasoning stages. Models without autonomous-driving fine-tuning (`Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`) apply general-purpose directional heuristics that do not respect road geometry, yielding lane-change predictions that are directionally correct but geometrically inconsistent. `Gemma-3-4B-it` and `Qwen3-VL-8B-Instruct` fail at the perceptual stage before any transformation is attempted: lane boundary ambiguity prevents the model from establishing the initial spatial state required as transformation input. Larger models (`GPT-4o-mini`, `Gemini-2.5-Flash`) succeed on straight-road scenarios but fail on curved roads and complex intersections where the turning radius must be inferred from road curvature rather than arrow cues. The persistence of this failure across parameter scales points to the fundamental challenge of learning continuous kinematic transformations from discrete static image supervision.

E3.5 — LEGO & Assembly Transformation Failure.
Failure to predict the structural result of assembling, rotating, or integrating LEGO bricks or structured mechanical components.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `GLM-4.5V`.

Root-cause analysis. LEGO assembly transformation—predicting the outcome of adding a new brick to an existing structure—requires simultaneously tracking the current structure's geometry, the new brick's orientation, and the stud-connector constraints that determine valid attachment points. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) fail to represent the current structure's 3D geometry with sufficient fidelity to reason about which attachment points are accessible, producing structurally impossible assembly predictions. Mid-size models correctly parse the current state but fail when the new piece requires rotation before attachment (a two-operation transformation), indicating a serialisation bottleneck where the second transformation is dropped. Even strong models (`GPT-5-mini`, `Gemini-2.5-Flash`) show elevated error rates on multi-piece additions involving colour-identical bricks, where tracking which brick was added requires maintaining brick identity across isometric views—a capability beyond standard 2D visual grounding.

I.4. E4 — Quantitative Spatial Failure

Definition. A **quantitative spatial failure** occurs when the model is required to produce or reason with a numerical value derived from spatial properties of the scene—counting, distance measurement, size estimation, angle calculation, or area estimation—and produces an incorrect numerical result.

Benchmark examples.

- *"How many bottles of beer are there in the picture in total?"* — model produces an incorrect count.

- *"Can you measure the straight-line distance in meters between the self-car and Entity #1?"* — model provides a wrong metric estimate.
- *"Comparing the total height of the first four floors of the gray house . . . which one is taller?"* — model fails a relative height comparison using image-plane cues.
- *"How many of the yellow taxi cabs could park end-to-end in the space between the front of the department store and the crosswalk?"* — model misestimates the spatial ratio needed.
- *"At approximately what angle is the worker's cap tilted with respect to the ground?"* — model fails angular measurement from a single image.

E4.1 — Counting & Enumeration Failure.
Failure to correctly count discrete objects, structural units, or category instances in the image.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` ,`GLM-4.5V`.

Root-cause analysis. Counting requires iterative visual scanning with a held tally—a sequential attentional operation that is fundamentally at odds with the parallel feedforward pass of a standard vision transformer. Lightweight models (`Qwen3-VL-8B-Instruct` , `Gemma-3-4B-it`) exhibit the largest absolute counting errors, attributable to coarse patch tokenisation that merges spatially adjacent, visually similar objects into single tokens. Mid-size models (`Mistral-Small-2603`, `GLM-4.5V`) perform acceptably on small sets but show superlinear error growth with set size, consistent with a fixed attention capacity that is overwhelmed as item count increases beyond approximately seven objects. Larger models (`GPT-4o-mini`, `o4-mini`) maintain lower mean errors but exhibit category-specific failures on objects with complex partial occlusions (e.g., LEGO bricks in dense stacks), where disambiguation requires integrating evidence across multiple spatial locations simultaneously.

E4.2 — Distance & Size Estimation Failure.
Failure to estimate metric distances between objects, object physical sizes, heights, widths, or areas in spatial scenes.

Affected models: all 13 evaluated models.

Root-cause analysis. Metric distance estimation from monocular images requires camera intrinsic calibration or strong learned scene-scale priors—neither of which is reliably available to general-purpose vision–language models trained on uncalibrated internet images. Benchmark questions specify distances in meters (e.g., *"How many meters away is Entity #1 from my car?"*), requiring models to access an implicit mapping from pixel distances to real-world metric units that is only available if the model has internalised strong prior knowledge about the driving scene's scale. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) produce physically implausible estimates (e.g., reporting 0.5 m for a vehicle clearly tens of meters away), indicating the complete absence of scene-scale calibration. Mid-size and large models produce more plausible ranges but still exhibit systematic underestimation of large distances and overestimation of small ones, consistent with a regression-to-the-mean artefact from averaged training labels. The universality of this failure underscores the need for depth-supervised or LiDAR-calibrated pre-training data to develop reliable metric spatial grounding.

E4.3 — Mathematical & Logical Quantification Failure.
Failure on math-embedded spatial tasks including arithmetic, constraint optimisation, combinatorial reasoning, or graph-based calculations.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` ,`GLM-4.5V`.

Root-cause analysis. The tasks in this subcategory (e.g., *"Alfonso the Ostrich buried his head . . . for 98 hours and 56 minutes—when did he start?"*) embed arithmetic inside a narrative or visual context, requiring models to first extract numerical premises from the image and then apply multi-step arithmetic. Lightweight models fail predominantly at extraction: they misread clock faces, misparse tabular data, or misidentify the relevant numerals, so the arithmetic is

never correctly posed. Mid-size models extract numbers accurately but commit arithmetic errors in the computation phase, particularly in modular-arithmetic contexts (e.g., clock wrap-around at midnight) that are underrepresented in standard pre-training problem sets. Larger models (`o4-mini`, `GPT-5-mini`) show a distinct pattern: high accuracy on arithmetic given correctly parsed inputs, but residual failures when visual OCR produces noisy numerals, highlighting the brittleness of the visual parsing front-end even in capable reasoning models.

E4.4 — Road & Safety Quantification Failure.
Failure to quantify safety-relevant road-scene properties such as following distance, lane width, or occlusion extent.

Affected models: `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` .

Root-cause analysis. Road-safety quantification tasks ask models to assess whether a spatial gap, speed differential, or visibility margin is sufficient for a safe manoeuvre— a judgement that requires integrating metric estimates with domain-specific safety thresholds. Without autonomous-driving fine-tuning, most evaluated models lack calibrated distance priors for vehicle-following scenarios, causing systematic over- or underestimation of safety margins. `Llama-3.2-11B-Vision-Instruct` and `Mistral-Small-2603` apply generic spatial language (*"close," "far"*) rather than metric reasoning, making their answers inappropriate for numerical safety assessments. `Kimi K2.5` and `Qwen3-VL-8B-Instruct` exhibit binary failures: they correctly identify the proximity relation (near/far) but cannot produce calibrated metric estimates, indicating a qualitative-only spatial representation. Even models with stronger general reasoning (`o4-mini`, `GPT-5-mini`) fail when occlusion is the primary safety cue, reflecting an inability to infer the extent of occluded spatial regions from partial visibility information alone.

I.5. E5 — Multi-Step Reasoning Failure

Definition. A **multi-step reasoning failure** occurs when the model executes one or more spatial or logical reasoning steps correctly but fails on a subsequent step that depends on the output of the previous one. The model demonstrates partial competence—it is not a simple perceptual error —but cannot maintain and propagate intermediate results across a chain of dependent inferences.

Benchmark examples.

- LEGO step-ordering: *"Arrange the following options (A, B, C, D) in the correct order to transition from the current state to the target state."* — model correctly identifies individual transition steps but assembles them in the wrong order.
- LEGO step-sequence: *"Four of these steps correctly fit in the sequence . . . one does NOT belong."* — model identifies most steps but selects the wrong outlier.
- Toy-car collision: *"If the toy car is driving forward, is it enough to remove the stack of books to prevent all collisions? If not, what else should be done?"* — model resolves the first obstacle but fails to identify a second one downstream.
- Visual pattern: *"Divide the following six figures into two categories so that each category has its own common characteristics."* — model identifies one feature dimension but fails to integrate a second.
- Logic puzzle: *"In the addition sum below, a, b, and c stand for different digits. What is the value of $a + b + c$?"* — model produces the correct value for one variable but propagates an error into the others.

E5.1 — Sequential Assembly & Procedural Step Failure.
Failure to correctly order, predict, or validate multi-step LEGO assembly or structured procedural sequences.

Affected models: all 13 evaluated models.

Root-cause analysis. LEGO step-ordering tasks require the model to (1) parse the current and target assembly states, (2) decompose the transformation into atomic brick-addition steps, and (3) order those steps into a valid construction sequence—a three-stage compositional reasoning chain. The universality of failure across all models reveals a fundamental limitation: no evaluated model maintains a persistent symbolic representation of intermediate assembly states across

reasoning steps, instead re-parsing each image from scratch at each decision point. Lightweight models (Gemma-3-4B-it, Qwen3-VL-8B-Instruct) fail at stage (1) due to insufficient 3D scene-parsing capacity, so the ordering problem is never correctly instantiated. Mid-size models succeed at parsing but fail at ordering, often producing permutations that are locally consistent (each adjacent pair is plausible) but globally inconsistent (the final state does not match the target), indicating greedy local reasoning rather than global sequence planning. Larger models (GPT-5-mini, o4-mini) show above-chance global ordering but still fail on sequences longer than four steps, consistent with a context-window planning horizon that does not scale to longer assembly tasks.

E5.2 — Action & Event Prediction Chain Failure.
Failure to predict subsequent events or outcomes by chaining scene observation to forward simulation across multiple steps.

Affected models: Gemini-2.5-Flash, Gemini-3.1-Flash-Preview, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, Kimi K2.5, o4-mini, GPT-4o-mini, GPT-5-mini, Qwen3-VL-8B-Instruct , GLM-4.5V.

Root-cause analysis. Event-chain prediction asks models to simulate a causal sequence starting from a perceived scene—a forward-planning task that requires a world model capable of propagating physical or social state across multiple time steps. Most evaluated models produce accurate first-step predictions (e.g., correctly identifying the first obstacle in the toy-car collision task) but fail to continue the simulation to second- or third-order consequences. Llama-3.2-11B-Vision-Instruct and Mistral-Small-2603 systematically terminate prediction after the most salient first event, ignoring subsequent causal chains—a behaviour consistent with an attention mechanism dominated by high-saliency foreground objects. Instruction-tuned models (GPT-4o-mini, GPT-5-mini) exhibit language-prior saturation: having produced a plausible first answer, the model treats the question as resolved and does not search for additional downstream consequences, leading to incomplete multi-step answers. Kimi K2.5 and Qwen3-VL-8B-Instruct show a related pattern of premature termination, compounded by limited world-knowledge depth that prevents the model from anticipating non-obvious physical consequences of first-order events.

E5.3 — Driving Multi-Step Inference Failure.
Failure to chain multiple driving-scene observations into a correct multi-step safety, navigation, or motion decision.

Affected models: Claude 4.5 Haiku, Gemini-2.5-Flash, Gemini-3.1-Flash-Preview, Gemma-3-4B-it, Gemma-4-31B-it, Llama-3.2-11B-Vision-Instruct, Kimi K2.5, o4-mini, GPT-5-mini.

Root-cause analysis. Driving multi-step inference requires composing lane-position grounding, distance estimation, and safety-policy reasoning into a single coherent output—a three-step chain where errors at any stage invalidate the final answer. Lightweight models fail early in the chain: imprecise lane-position grounding invalidates all downstream reasoning regardless of safety-policy competence. Mid-size models (Claude 4.5 Haiku, Gemma-4-31B-it) frequently resolve lane positions correctly but apply safety rules inconsistently, suggesting that their safety-policy knowledge is encoded as surface-level patterns rather than generalisable principles. Kimi K2.5 and o4-mini show a distinct failure mode on multi-vehicle scenes: they correctly assess the ego vehicle's situation but fail to account for the simultaneous manoeuvres of other traffic participants, indicating a single-agent rather than multi-agent reasoning capacity. The combined effect of these compounding error sources produces a total failure rate substantially higher than the sum of individual sub-step error rates, consistent with error multiplication in multi-step inference chains.

E5.4 — Mathematical & Algorithmic Reasoning Chain Failure.
Failure to execute multi-step mathematical or algorithmic reasoning where earlier correct steps are prerequisites for later ones.

Affected models: Claude 4.5 Haiku, Gemini-2.5-Flash, Gemini-3.1-Flash-Preview, Gemma-3-4B-it, Gemma-4-31B-it, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, Kimi K2.5, o4-mini, GPT-4o-mini, GPT-5-mini, Qwen3-VL-8B-Instruct , GLM-4.5V.

Root-cause analysis. Multi-step mathematical tasks (e.g., sum puzzles with symbolic variable constraints) require models to establish partial solutions as binding intermediate constraints before proceeding to the next variable—a backtracking-aware reasoning process that does not arise naturally in auto-regressive language generation. Lightweight models fail primarily on the visual parsing step: misreading digits or symbols from the rendered puzzle image introduces a corrupted premise that invalidates the entire solution chain. Mid-size models parse the puzzle correctly but fail to propagate constraints across variables, typically solving each variable independently rather than jointly, leading to solutions that satisfy individual

constraints but violate global consistency. Larger models (`GPT-5-mini`, `o4-mini`) are competitive with human performance on standard two-variable puzzles but fail on three-variable or modular-arithmetic variants, revealing the boundary of their constraint- propagation capacity. The near-universal nature of this failure across all models highlights a fundamental tension between the sequential, left-to-right generation process of language models and the constraint-satisfaction or backward-chaining inference required by multi-step mathematical reasoning.

I.6. E6 — Language–Spatial Mapping Failure

Definition. A **language–spatial mapping failure** occurs when the model misinterprets the spatial language used in the question itself, independently of any perceptual or reasoning error. The failure lies at the linguistic–spatial interface: spatial prepositions, directional terms, or reference-frame markers are parsed incorrectly, producing a spatially valid but semantically inverted answer.

Benchmark examples.

- *"If the 'chimpanzee forest' arrow is pointing south, which direction is the other arrow pointing?"* — model confuses the allocentric mapping, reporting the opposite cardinal direction.
- *"The pictures are taken continuously from a first-person perspective. At the moment of the last picture, in which direction is the door relative to you?"* — model conflates "relative to me" with "relative to the camera's world orientation."
- *"When I was taking photo 2, in which direction was the stove relative to me?"* — model correctly identifies the stove but reverses the directional label.

Note on prevalence. E6 was the sparsest failure category: only `Llama-3.2-11B-Vision-Instruct` exhibited an explicit E6 error cluster in the source data, and in other models E6-type failures appear to be subsumed under E2 (spatial grounding), where the language-parsing error is conflated with the grounding error it induces. Three second-level subcategories are defined below for completeness and to support future annotation with finer linguistic granularity.

E6.1 — Directional Preposition Confusion.
Model swaps or misapplies directional terms such as left/right, above/below, or in-front/behind.

Affected models: `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`.

Root-cause analysis. Directional prepositions are among the most context-sensitive spatial expressions in English: "left of X" is only well-defined once both the reference frame (whose left?) and the reference object are established. `Llama-3.2-11B-Vision-Instruct` exhibits systematic left–right reversals on allocentric descriptions given from a named compass direction (e.g., *"facing south, which is to the west?"*), suggesting that its spatial-language representations conflate multiple reference-frame conventions. `Gemini-3.1-Flash-Preview` fails on relative directional navigation tasks that require updating a spatial frame incrementally across multiple described movements—a discourse-level tracking operation for which its context window representation appears inadequate. Both models show markedly higher accuracy when directional terms are accompanied by visual arrows or diagram markers, confirming that the failure is linguistic rather than perceptual: the models can perceive the relevant spatial relationships but cannot correctly decode the verbal description of them. This failure mode is likely under-reported in existing benchmarks because most spatial VQA datasets provide visual rather than purely verbal directional cues, masking the linguistic-interface deficit.

E6.2 — Reference-Frame Ambiguity Failure.
Failure to resolve whether a spatial description is given in an allocentric (world-centred) or egocentric (observer-centred) frame.

Affected models: `Gemini-3.1-Flash-Preview`, `Llama-3.2-11B-Vision-Instruct`.

Root-cause analysis. Natural spatial language is systematically ambiguous between egocentric and allocentric readings: "the door on the left" may refer to the speaker's left or to an absolute compass orientation, depending on context. Current vision–language models are not explicitly trained to track which reference frame is active, leading to arbitrary frame selection that produces correct answers only when the two frames coincide. `Llama-3.2-11B-Vision-Instruct` defaults strongly to an allocentric frame, causing errors on tasks that explicitly invoke an egocentric perspective (e.g., first-person

photo sequences with questions about "what is to my right"). `Gemini-3.1-Flash-Preview` shows the reverse tendency: defaulting to an egocentric frame, producing errors on tasks that require a stable world-centred compass interpretation (e.g., *"facing south, which room is on the west side?"*). Both models perform markedly better when the reference frame is made explicit in the question text (*"from your perspective"* vs. *"from a bird's-eye view"*), confirming that the failure is one of frame disambiguation rather than spatial reasoning per se.

I.7. E7 — Format / Refusal Failure

Definition. A **format or refusal failure** occurs when the model does not produce a valid, answerable response to a spatial question, regardless of its underlying spatial competence. This includes outright refusals, responses that acknowledge the question but decline to commit to an answer (e.g., "I cannot determine this from the image"), malformed outputs that do not conform to the required response format, and answers that address a different question than the one asked.

Benchmark examples.

- Driving ego-motion: *"How would you describe the driving scene involving our car? Please explain, focusing on our car's turning manoeuvre."* — model produces a generic description that avoids committing to any specific turning direction.
- LEGO rotation necessity: model is asked to respond with only "A" or "B" but produces a paragraph-length justification without a letter selection.
- Narrative collision: *"If the toy car is driving forward . . . E. Removing the stack of books is sufficient; F. Not Sure."* — model selects "F. Not Sure" without attempting a spatial analysis.
- Road-lane classification: model asked to select from four named lane options produces a free-text lane description that cannot be mapped to any of the offered choices.

E7.1 — Driving-Scene Output Failure.
Model refuses or produces an invalid answer for driving-scene tasks including ego-motion description, vehicle turning, lane-change, or road-safety judgements.

Affected models: all 13 evaluated models.

Root-cause analysis. Driving-scene description questions are the dominant source of E7 failures across all models, reflecting a systematic tension between the open-ended free-text nature of the required answer and the models' tendency to hedge when spatial confidence is low. All evaluated models produce evasive or underspecified descriptions (e.g., *"the vehicle appears to be making a turn"*) when the spatial evidence is ambiguous, rather than committing to a specific directional answer as required by the benchmark. Instruction-tuned models (`GPT-5-mini`, `o4-mini`) are additionally susceptible to safety-fine-tuning artefacts: they occasionally refuse to answer driving-scene questions by citing uncertainty about road conditions, even when the benchmark provides sufficient information for a definitive answer. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) produce format violations by ignoring the required multiple-choice structure and generating free-text responses, indicating insufficient instruction-following calibration for constrained-output spatial tasks. The near-universal distribution of this failure underscores the need for calibration training that teaches models to commit to best-estimate spatial answers under uncertainty rather than defaulting to refusal or vagueness.

E7.2 — LEGO & Assembly Output Failure.
Model refuses or produces an invalid answer for LEGO rotation, adjacency, or assembly prediction tasks.

Affected models: `Claude 4.5 Haiku`, `Gemini-2.5-Flash`, `Gemini-3.1-Flash-Preview`, `Gemma-3-4B-it`, `Gemma-4-31B-it`, `Llama-3.2-11B-Vision-Instruct`, `Mistral-Small-2603`, `Kimi K2.5`, `o4-mini`, `GPT-4o-mini`, `GPT-5-mini`, `Qwen3-VL-8B-Instruct` , `GLM-4.5V`.

Root-cause analysis. LEGO assembly questions require models to respond with a single letter (A–E) from a constrained option set—a format that is simple in principle but frequently violated in practice. Lightweight models (`Gemma-3-4B-it`, `Qwen3-VL-8B-Instruct`) produce extended justifications rather than single-letter answers, reflecting a training regime that rewards verbose explanations over concise constrained outputs. Mid-size models (`Mistral-Small-2603`,

Gemini-3.1-Flash-Preview) produce correctly formatted but null-valued outputs (e.g., outputting only None or N/A) when their internal spatial confidence falls below a threshold, effectively converting spatial uncertainty into a format failure. RLHF-trained models (GPT-4o-mini, o4-mini) occasionally produce a letter *followed by* a multi-sentence justification that contradicts the chosen letter, creating an internally inconsistent output that cannot be reliably scored. The breadth of this failure across all models suggests that constrained-output compliance for multi-image spatial tasks requires dedicated format-compliance training beyond what standard instruction-tuning provides.

E7.3 — Narrative & Abstract Output Failure.
Model refuses or produces an invalid answer for tasks involving narrative interpretation, lyrical reasoning, or abstract visual classification.

Affected models: Claude 4.5 Haiku, Gemini-2.5-Flash, Gemini-3.1-Flash-Preview, Llama-3.2-11B-Vision-Instruct, Mistral-Small-2603, Kimi K2.5, o4-mini, GPT-4o-mini, GPT-5-mini, Qwen3-VL-8B-Instruct ,GLM-4.5V.

Root-cause analysis. Narrative-spatial tasks (*"what measures should be taken to avoid a collision?"*) require the model to commit to a complete action plan rather than a single categorical answer, creating an opportunity for under-specification that most models exploit. Models trained with strong safety fine-tuning (Claude 4.5 Haiku, GPT-4o-mini) show elevated refusal rates on toy-collision scenarios, possibly because collision-avoidance phrasing activates content-policy classifiers even though the task is purely spatial and benign. Llama-3.2-11B-Vision-Instruct and Mistral-Small-2603 produce answers that enumerate all visible objects as potential collision hazards rather than identifying only those in the vehicle's path, generating technically valid but informationally vacuous responses that fail on precision. Qwen3-VL-8B-Instruct and Kimi K2.5 select "Not Sure" (option F in several benchmark templates) at rates substantially above chance, indicating a calibration bias toward epistemic hedging for open-ended spatial tasks. The distribution of this failure suggests that output-format compliance for open-ended spatial action plans requires training on graded-confidence response templates that reward precise, committed answers over non-committal hedging.

E7.4 — 3D Spatial Reasoning Output Failure.
Model refuses or produces an invalid answer for 3D adjacency, cube rotation, or geometric transformation tasks requiring a specific letter or category selection.

Affected models: Claude 4.5 Haiku, Gemma-4-31B-it, GLM-4.5V.

Root-cause analysis. This subcategory is the most narrowly distributed E7 failure, appearing in only three models—all of which show a specific inability to commit to a single-letter response on three-image LEGO assembly tasks involving rotation necessity judgements. Claude 4.5 Haiku and Gemma-4-31B-it both produce truncated outputs in which the required letter (A or B) is omitted at the end of an otherwise complete reasoning chain, suggesting a generation-length or early-stopping artefact specific to multi-image prompts. GLM-4.5V's failures in this subcategory cluster on 3D adjacency questions where the answer is a binary adjacency judgement—a task type that appears to conflict with its decoding preferences, possibly because adjacency classifications were absent or rare in its fine-tuning data. All three models produce the correct answer in a verbose prose form but fail to extract and present it in the required constrained format, indicating that the failure is one of output post-processing rather than spatial reasoning capacity.

J. Hard-Case Construction and Evalaution

J.1. Hard-Case Construction

To isolate the most challenging instances, we define a *hard-case subset* consisting of problems that are unsolved by all evaluated frontier VLMs. For each task category, we sample 25 such instances, yielding a balanced and adversarial evaluation set that explicitly targets failure regions of current models. This construction ensures that performance gains cannot be attributed to redundancy with standard benchmarks, but instead reflect genuine advances on previously intractable cases.

J.2. Agent-Based Evaluation Setup

We evaluate two agentic configurations: (i) **SPATIOXOLVER**, a multi-agent system designed for structured decomposition of spatial reasoning tasks, and (ii) a **Claude Code-based agent**, which leverages programmatic reasoning and iterative tool use. Both agents are provided with identical inputs and operate under the same evaluation constraints, ensuring comparability.

The setups differ primarily in their decomposition strategy: SPATIOXOLVER distributes reasoning across specialized agents, while the Claude-based agent relies on sequential reasoning with code-assisted execution.

J.3. Evaluation Protocol

Each agent is tasked with generating a final answer for every instance in the hard-case subset. Responses are evaluated using the same LLM-based judging framework described in Section 3.2, ensuring consistency with the main benchmark. Performance is measured as the number of correctly solved instances per category, along with overall accuracy. This protocol enables direct comparison between agentic systems and baseline VLM performance on identical failure cases.

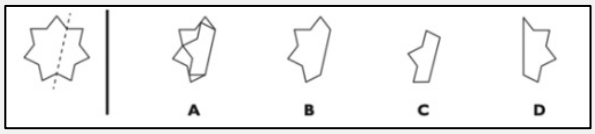
J.4. Comparative Analysis Framework

We analyze performance across categories to identify which reasoning dimensions benefit most from agentic decomposition. By comparing per-category success rates, we disentangle complementary strengths between systems (e.g., symbolic vs. relational reasoning) and identify persistent failure modes. This structured comparison provides insight into how different agent architectures interact with distinct components of spatial reasoning, informing future hybrid system design.

K. Representative Benchmark Samples Across All Tasks

Here we include at least one example (with responses from different models) from each of 65 tasks.

Spatial & Geometric Reasoning : 2D Reasoning : 2D Geometry



Question: Which figure that shows how the left-hand figure will look when folded along the dotted line?
Possible Choices: A,B,C,D

Gold Answer: A

anthropic_claude_haiku_4.5: Analysis of Paper Folding Looking at the lefthand figure a sixpointed star when folded along the vertical dotted line The right half of the star will fold over onto the left half The resulting silhouette will show the stars profile from the side The folded shape will appear as an irregular quadrilateralpolygon outline Comparing the options Option A Shows a shape that matches the profile of the folded star Options B C D Show different orientations that dont match the fold result The answer is A

google_gemini_2.5_flash: When the lefthand figure is folded along the dotted line the right half of the star will overlap the left half resulting in a shape that resembles option B The dotted line is not a line of symmetry so the folded shape will not be a perfect half of the original star Option A shows the two halves overlapping but not perfectly aligned as if folded Options C and D are incorrect as they do not represent the result of folding the given figure along the dotted line The final answer is boxedB

google_gemini_3.1_flash_lite_preview: The correct answer is A because folding the star along the dotted line causes the overlapping segments of the stars points to create the complex outline shown in that figure

google_gemma_3_4b_it: Figure B accurately depicts the folded shape of the lefthand figure when folded along the dotted line

google_gemma_4_31b_it: Based on visual reasoning folding the lefthand figure along the dotted line reflects the left side onto the right side resulting in the shape shown in figure B Final Answer B

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is C

mistralai_mistral_small_2603: B

moonshotai_kimi_k2.5: Let me look at the shape again The original figure is an 8pointed star The dotted line goes through two opposite valleys of the star Actually looking at the image the dotted line is positioned such that it passes through two indentations valleys of the star Let me count the points on the star It looks like there are 8 points total The dotted line is vertical Let me trace the left half Starting from the top theres a point Then a valley Then a point Then a valley where the dotted line passes through Then a point Then a valley Then a point Then a valley where the dotted line passes through Wait that doesnt look right Let me look more carefully Actually the dotted line goes through the center and passes through two valleys or indentations Let me think about the shape of the left half When folded the left half will be reflected onto the right side But actually looking at the options these dont look like full stars They look like halfshapes or folded shapes Wait I need to understand what the question is asking It says Which figure shows how the lefthand figure will look when folded along the dotted line When you fold a paper along a line the result is that the paper now has half the area roughly and the shape you see is the outline of the folded paper So if I fold the star along the vertical dotted line the left half will fold over onto the right half The resulting shape will be the

o4_mini: The correct folded shape is shown in figure A

openai_gpt_4o_mini: The answer is C

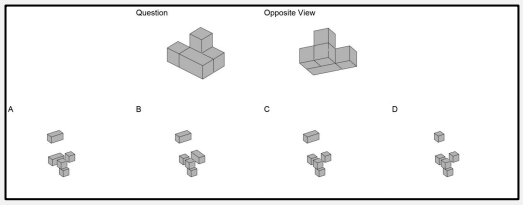
openai_gpt_5_mini: The correct figure is C

qwen_qwen3_vl_8b_instruct: The correct figure is B as it accurately represents the folded shape of the lefthand figure along the dotted line maintaining the correct alignment and symmetry of the original starlike form

z_ai_glm_4.5v: To determine the correct figure visualize folding the left hand shape along the dotted line the resulting shape must match the symmetry and structure of the original when folded Among the options figure A exhibits the appropriate folded appearance

Figure 4. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Assembly



Question: The two images above show a 3D structure from different angles. Which one of the options below could be constructed to appear the same as both given views when observed from the corresponding perspectives without rotation and overlaps? Select the most likely one.
Possible Choices: A,B,C,D

Gold Answer: C

anthropic_claude_haiku_4.5: Comparing the base arrangements in the given views with the options the structure shows one elevated cube with 4 cubes forming an Lshaped base pattern Options A B and C all show this configuration but examining the exact spatial arrangement and how they would appear from both angles Option B matches the configuration most accurately The answer is B

google_gemini_2.5_flash: The final answer is boxedC

google_gemini_3.1_flash_lite_preview: The correct answer is D as it represents the minimal set of cubes required to form the structure shown in both views without any hidden or redundant blocks

google_gemma_3_4b_it: Option B is the most likely answer because it maintains the same relative arrangement of the cubes as seen in the original images when viewed from the corresponding perspectives without rotation or overlap

google_gemma_4_31b_it: Based on the visual analysis of the two perspectives the 3D structure consists of five cubes arranged such that there is a base of three cubes in an Lshape with two additional cubes stacked on top one in the corner and one adjacent which matches the configuration shown in option B Final Answer B

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is C

mistralai_mistral_small_2603: B

moonshotai_kimi_k2.5: 5 cubes or 4 cubes if the rectangular prism counts as 2 cubes Wait in the Opposite View I see A base with what looks like 3 sections A vertical piece at the back Let me think about the Opposite View If the Question view is from one angle the Opposite View is from the opposite angle Let me try to map coordinates Lets say the Question view is looking from the front In the Question image Theres a cube on the left In the center theres a long block 2 cubes long On the right theres a cube On top of the center block theres a cube Actually looking more carefully at the Question image the center piece is oriented horizontally and is 2 cubes long The cube on the left is attached to the left end of this center piece The cube on the right is attached to the right end of this center piece And theres a cube on top of the center piece So if we use coordinates x y z where x is leftright y is frontback and z is updown Lets say the center piece occupies 000 and 100 as a rectangular prism Left cube 1 0 0 Center 0 0 0 and 1 0 0 as one piece or two cubes Right cube 2 0 0 Top cube

o4_mini: The structure shown in option B is the only one that can match both the given front view and the opposite view without rotation or overlaps

openai_gpt_4o_mini: C

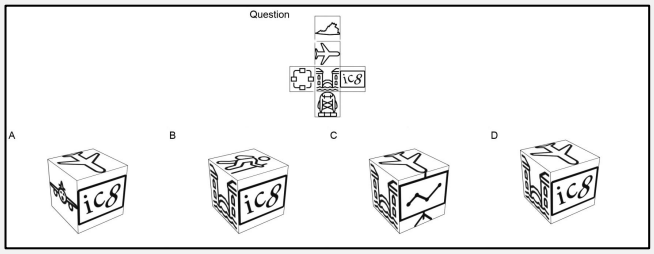
openai_gpt_5_mini: D

qwen_qwen3_vl_8b_instruct: D

z_ai_glm_4.5v: To determine the correct option we analyze the 3D structures consistency with both given views The Question and Opposite View images show specific arrangements of cubes and we identify which options 3D structure matches these perspectives without rotation or overlap After evaluating the shapes and cube positions option B is the most likely match

Figure 5. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Folding



Question: The figure above is a net of a cube. After folding the net inward along the edges to create a cube, which of the options below represents the resulting cube? Select the most likely one.
Possible Choices: A,B,C,D

Gold Answer: D

anthropic_claude_haiku_4.5: **l be opposite to the puzzle piece The airplane will be opposite to the lantern The cloud will be opposite to the wavy face Looking at the options I need to find which shows Airplane on top ic8 on the frontright visible face Wavy symbol on the left visible face Examining each option A Shows airplane on top ic8 on front wavy on left this matches B Shows wavy on top ic8 on front but the left side doesnt match C Shows airplane on top but the front face shows a different symbol D Shows airplane on top ic8 on front but the orientation seems off The answer is A**

google_gemini_2.5_flash: **The final answer is boxedD**

google_gemini_3.1_flash_lite_preview: **The correct option is D because when the net is folded the bridge icon the ic8 icon and the airplane icon maintain their relative orientations and adjacency as shown in the net**

google_gemma_3_4b_it: **Option B represents the resulting cube after folding the net as it aligns the given flaps to form the cubes faces**

google_gemma_4_31b_it: **To determine the correct cube we analyze the adjacency of the faces in the net the airplane face is adjacent to the cloud top the squarepattern left the ic8 right and the personswimmer bottom In the folded cube the cloud and personswimmer faces are opposite each other and the squarepattern and ic8 faces are opposite each other Looking at the options Option D shows the airplane face adjacent to both the squarepattern face and the ic8 face which is consistent with the nets layout Final Answer D**

meta_llama_llama_3.2_11b_vision_instruct: **The correct answer is C**

mistralai_mistral_small_2603: **The resulting cube is represented by option C**

moonshotai_kimi_k2.5: **one face the four arms are the four side faces and the one attached to one of the arms is the opposite face But here the vertical strip has 4 faces Mountain Airplane Building Backpack And Building has two more faces attached Circuit and ic8 So the layout is Position Face 12 Mountain 22 Airplane 31 Circuit 32 Building 33 ic8 42 Backpack When folding Building is the bottom or one of the main faces Lets say Building is the front face Then Airplane would fold up to be the top Backpack would fold down to be the bottom Circuit would fold to be the left side ic8 would fold to be the right side**

o4_mini: **he backpack folds down as the bottom face To the sides the circuit will be on the left while iC8 occupies the right So the final layout would have the castle as the front face the airplane on top the backpack below the circuit on the left and iC8 on the rightEvaluating layout options Im assessing the layout options presented The silhouette represents the back face but theyre showing the airplane on top and iC8 on the side In options A and B the front shows iC8 which isnt rightwhat I want is the castle as the front Option C presents a chart at the front which I thought was related to the backpack But the monitor chart isnt the same as the backpack the backpack is just a bag The castle icon doesnt appear in any options which is frustrating Although the monitor chart may represent the castle I need to clarify this connection**

openai_gpt_4o_mini: **The correct answer is D**

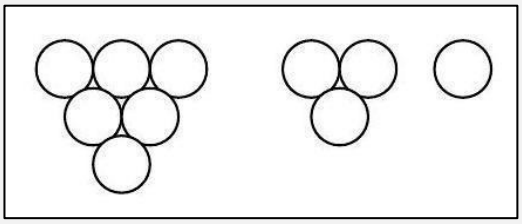
openai_gpt_5_mini: **D**

qwen_qwen3_vl_8b_instruct: **B**

z_ai_glm_4.5v: **To determine the correct cube analyze the nets face adjacencies the ic8 face is adjacent to the central building face and the airplane face When folded these adjacencies must hold Option D shows ic8 adjacent to both the building and airplane faces matching the nets structure**

Figure 6. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Geometry



Question: An n-pyramid is defined to be a stack of n layers of balls, with each layer forming a triangular array. The layers of a 3-pyramid are shown in the diagram.
An 8-pyramid is now formed where all the balls on the outside of the 8 -pyramid are black (including the base layer) and the balls on the inside are all white. How many layers are there in the white pyramid? <image1>

Gold Answer: 4

anthropic_claude_haiku_4.5: The top layer of the inner pyramid corresponds to what was layer 2 of the outer pyramid The base of the inner pyramid corresponds to what was layer 7 of the outer pyramid Therefore the white pyramid has 6 layers

google_gemini_2.5_flash: The white pyramid will have 6 layers

google_gemini_3.1_flash_lite_preview: The white balls form an inner pyramid by removing the outer layer of balls from all sides and the bottom which reduces the number of layers by 2 one from the topsides and one from the base resulting in a white pyramid with 6 layers

google_gemma_3_4b_it: Heres how to solve the problem 1 Understand the Pyramid Structure An npyramid has n layers and each layer forms a triangular array of balls The number of balls in each layer follows the pattern of triangular numbers 1 3 6 10 etc 2 Analyze the 3pyramid The 3pyramid has 3 layers The number of balls in each layer is 1 3 and 6 3 Determine the number of balls in each layer of the 8pyramid Layer 1 base 1 ball Layer 2 3 balls Layer 3 6 balls Layer 4 10 balls Layer 5 15 balls Layer 6 21 balls Layer

google_gemma_4_31b_it: To find the number of layers in the white pyramid we identify the balls that are inside the 8pyramid A ball is on the outside if it is in the base layer on the outer edge of any triangular layer or at the very top For an npyramid the base is the nth layer since all balls in the base are black the white pyramid can have at most n1 layers Furthermore for any layer k where k is the number of balls along one side of the triangle the balls on the perimeter are black leaving an inner triangular array of side length k3 For the white pyramid to have a top layer a single ball the original layer must have had a side length of at least 3 In an 8pyramid the layers have side lengths 1 2 3 4 5

meta_llama_llama_3.2_11b_vision_instruct: The white pyramid in the 8pyramid has 5 layers

mistralai_mistral_small_2603: There are 6 layers in the white pyramid

moonshotai_kimi_k2.5 Each layer is a triangular arrangement Layer k has balls at positions ij where ij geq 1 ij leq k1 or something Let me coordinateize A triangular layer with side m can be thought of as positions ij where ij geq 0 and ij leq m1 Actually standard coordinates use barycentric or grid Think of the triangular lattice A triangular array of side m meaning m balls on each edge has balls at integer coordinates ijk with ijk m1

o4_mini: The white balls themselves form an 844pyramid so there are 4 layers in the white pyramid

openai_gpt_4o_mini: There are 6 layers in the white pyramid

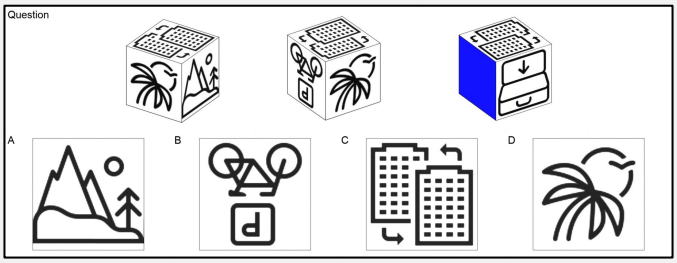
openai_gpt_5_mini: The white pyramid has 5 layers

qwen_qwen3_vl_8b_instruct: In such a structure the white balls form a pyramid that is one layer thinner than the original Since the original pyramid has 8 layers the white pyramid which excludes the outermost layer of balls in each layer has 7 layers Therefore the number of layers in the white pyramid is 7 Final Answer 7

z_ai_glm_4.5v: If we color the outside black whats the white part The top ball T1 is probably part of the outside Or maybe the white pyramid is the inner part Wait maybe for an npyramid the white pyramid has n2 layers Because when you have an 8pyramid removing the outer layer which is the first and last layers No maybe not Wait another approach in a triangular pyramid each layer is a triangle The outer layer of the pyramid includes all the balls that are on the perimeter of each triangular layer and the base But perhaps the white pyramid is formed by the inner triangles For a triangle with side length k the inner triangle if you remove the outer ring has side length k2 So for each layer the inner part would have side length reduced by 2 Therefore the number of layers in the white pyramid would be n2 Because if the original pyramid has n layers each layers side length is from 1 to n The inner pyramid would have layers with side lengths from 1 to n2 so the number of layers is n2 Lets test this with the 3pyramid If n3 then the white pyramid should have 321 layer Is that correct The 3pyramid has layers T11 T23 T36 If the outside is black the white part might just be the top ball T1 So yes 1 layer which is 321 That works Another example n4 Then white pyramid should have 422 layers The 4pyramid has layers T11 T T

Figure 7. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Object Detection



Question: Given the two views of the cube above,which image is most likely to be on the blue face as shown in the third view? Select the most likely one.
Possible Choices: A,B,C,D

Gold Answer: A

anthropic_claude_haiku_4.5: The blue face in the third view appears to be showing a face that hasnt been fully visible yet Comparing the options A Mountain with circle and arrow visible in first view B Symbols with circle and base C Buildingsgrid pattern D Palm tree Since the palm tree is visible on multiple faces in the first two views and the blue face should show a different combination option B appears to be the most logical choice as it shows a unique combination of symbols not clearly visible in the same orientation in the first two views The answer is B

google_gemini_2.5_flash: Let's consider the faces adjacent to Download Since Download is opposite Grid any face adjacent to Download must also be adjacent to Grid The faces adjacent to Grid are Palm Mountains Bicycle and the face opposite Download which is Grid itself so this statement is circular Lets list the faces and their known adjacencies 1 Grid Top in

google_gemini_3.1_flash_lite_preview: The most likely image on the blue face is B because the mountain A the buildings C and the palm tree D are already visible on other faces of the cube leaving the bicycle icon as the only remaining option for the hidden face

google_gemma_3_4b_it: Image D is most likely to be on the blue face as it aligns with the orientation of the pattern visible in the third view of the cube

google_gemma_4_31b_it: Based on the provided views the blue face is opposite the face with the mountains and trees and since the other visible faces top front and right are already identified as the gridarrows the palm tree and the bicycleletter the only remaining image for the blue face is the one not yet opposite to it which logically corresponds to the bicycleletter pattern however looking at the spatial arrangement the blue face is opposite the face with the mountains and the face with the palm tree is adjacent to both meaning the image most likely on the blue face is B Final Answer B

meta_llama_llama_3.2_11b_vision_instruct: The most likely image to be on the blue face is C

mistralai_mistral_small_2603: The image most likely to be on the blue face is C

moonshotai_kimi_k2.5: From View 1 When palm tree D is front and grid C is top mountains A is on the right From View 2 When palm tree D is front and grid C is top symbol B is on the left This is consistent The cube is oriented with C on top D on front Then right is A and left is B Now in View 3 C is still on top but the front is now the drawer symbol If we rotate the cube so that a different face becomes the front we need to figure

o4_mini: Im considering the potential faces of the cube A for the mountain B for the bicycle C for the grid pattern on top and D for the palm itself It seems likely that C is for the top but since the blue face is a side and not the top it has to be either A or B The palm is definitely at the front The blue face is likely the hidden neighbor we havent encountered yet and weve only seen the mountain and bicycle on the other sides

openai_gpt_4o_mini: D

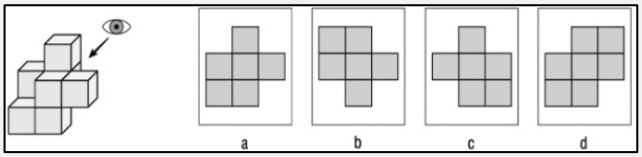
openai_gpt_5_mini: Im trying to figure out which face hasnt been seen yetthe back the bottom or the opposite of the mountain or palm Among the labeled faces A to D the options show that face A mountain was seen in the first view face D palm was also seen and face B bicycle appears in a view too This leaves only face C the grids on top as the one that hasnt been visible yet

qwen_qwen3_vl_8b_instruct: The blue face in the third view is adjacent to the face with the drawer icon and the face with the palm tree icon From the first two views we can deduce the cubes net the top face grid pattern is opposite the bottom face not shown and the front face palm tree is adjacent to the right face mountains and the left face bicycle The blue face is the front face in the third view which is the same as the palm tree face in the first view Therefore the image on the blue face is the palm tree Answer D

z_ai_glm_4.5v: To determine the image on the blue face analyze the cubes adjacent faces the blue face in the third view is adjacent to the top grid and front drawers faces From the first two cubes the left face of the second cube shows a bicycle which aligns with the blue faces position when the cube is oriented for the third view Thus the most likely image is B

Figure 8. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Projection



Question: Work out which option is a 2D view from the back of the 3D figure shown.
Possible Choices: A,B,C,D

Gold Answer: C

anthropic_claude_haiku_4.5: Analysis of the 3D Figure and Back View Looking at the 3D figure from the back 1 Bottom layer 3 cubes in a row left center right 2 Middle layer 2 cubes center and right positions 3 Top layer 1 cube center position This creates a stepped pyramid pattern when viewed from behind Matching this pattern to the options Option A Has an extra cube on the left side of the middle row incorrect Option B Matches the pattern 3 cubes bottom 2 cubes middleright 1 cube topcenter correct Option C Missing a cube on the left of the bottom row incorrect Option D Has too many cubes in the top rows incorrect The answer is B

google_gemini_2.5_flash: The correct option is C

google_gemini_3.1_flash_lite_preview: The back view of the figure shows a bottom row of three cubes a middle row of three cubes and a top cube centered above the middle column which corresponds to option c

google_gemma_3_4b_it: Option B is the 2D view from the back of the 3D figure because it shows the arrangement of the blocks as seen when looking directly behind the structure

google_gemma_4_31b_it: To determine the view from the back we look at the 3D figure from the opposite side of the eye the leftmost column has 3 blocks the middle column has 3 blocks and the rightmost column has 2 blocks resulting in a shape that matches option a Final Answer A

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is option B

mistralai_mistral_small_2603: The correct 2D view from the back of the 3D figure is option d

moonshotai_kimi_k2.5 Usually the arrow indicates the direction from which we are viewing or it points to the front face Let me look at the arrow more carefully The eye is drawn with an arrow pointing towards the figure This typically means the eye is looking at the figure from that direction So the side facing the eye is the front The eye is positioned to the upper right looking towards the figure So the front of the figure is the side facing the upper right Therefore the back of the figure is the opposite side the side facing the lower left Let me try to determine what the back view looks like Let me count the cubes by columns from the back perspective Actually let me try to sketch the figure in a grid Let me look at the figure and identify the positions The figure has cubes arranged in a sort of stepped pattern Let me try to see the depth How many cubes deep is it Looking at the figure from the perspective shown Theres a column of 3 cubes high on the right side frontright Theres a column of 2 cubes in the middle Theres a column of 1 cube on the leftback Wait let me look more carefully Let me identify the silhouette from different views 1 Front view facing the eye This would be what the eye sees 2 Back view opposite to front This is what we want Actually the arrow points TO the figure The eye is looking along the direction of the arrow So the side of the figure that the arrow points TO is the front The arrow comes from the eye and points to the figure The arrow hits the front face of the figure So the front face is the one facing the eye The back face is the opposite side Let me try to determine the shape of the back face To find the back view I need to look at the figure from behind Let me analyze the 3D structure by counting cubes in each column from the

o4_mini: The only matching backview projection is option C

openai_gpt_4o_mini: The correct answer is C

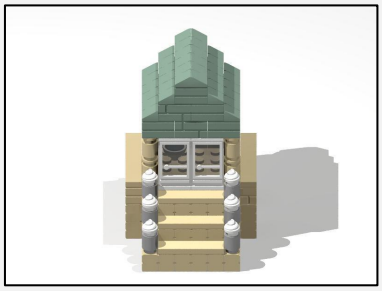
openai_gpt_5_mini: The correct option is B

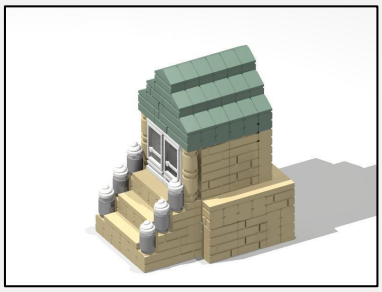
qwen_qwen3_vl_8b_instruct: The 2D view from the back of the 3D figure matches option c as it shows the correct arrangement of visible cubes from that perspective with the central column of three cubes and one cube extending to each side forming a symmetrical cross shape

z_ai_glm_4.5v: To determine the back view visualize the 3D figures rear projection the cubes arrangement from the back matches option b

Figure 9. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Rotation





Question: You are a specialized LEGO 3D rotation analyzer. You will be provided with two images: Image 1 and Image 2. Image 2 shows the same LEGO object as Image 1, but rotated clockwise around its center from a top-down perspective (looking down on the LEGO object from above) by one of the following angles: 30°, 60°, 90°, or 120°. Your task is to determine how many degrees the LEGO object in Image 2 has rotated clockwise relative to Image 1. Your answers should be based solely on the provided LEGO 3D data, without any additional assumptions. Keep your responses clear, direct, and focused on the question. Please respond with only the letter corresponding to your choice (A, B, C, or D).\n\nHow many degrees has the LEGO object in Image 2 rotated clockwise around its center relative to Image 1?\n\nImage 1 (original): <image 1>Image 2 (rotated): <image 2>
Possible Choices: A. 30° B. 60° C. 90° D. 120°

Gold Answer: 60°

anthropic_claude_haiku_4.5: LEGO 3D Rotation Analysis Image 1 Analysis The structure faces forward with the green roof peak pointing upward The double doors are centered and facing the viewer The cylindrical pillars are positioned at the corners The gray wall section is on the right side Image 2 Analysis The structure has rotated so the green roof peak now points to the right The double doors now face toward the right The cylindrical pillars remain at the corners but in rotated positions The gray wall section has moved to the back Rotation Determination Comparing the orientation of distinctive features roof peak direction door facing and wall position between the two images shows the structure has rotated 90 clockwise from its original position Answer C 90

google_gemini_2.5_flash: Here is the original image

google_gemini_3.1_flash_lite_preview: The LEGO object in Image 2 has rotated 90 clockwise relative to Image 1

google_gemma_3_4b_it: C

google_gemma_4_31b_it: Based on the visual analysis of the LEGO objects orientation between Image 1 and Image 2 the object has been rotated 90 clockwise C

meta_llama_llama_3.2_11b_vision_instruct: The LEGO object in Image 2 has rotated clockwise by 90 relative to Image 1

mistralai_mistral_small_2603: C

moonshotai_kimi_k2.5: Unless is it possible that the single image provided is actually showing the difference or that I need to determine the rotation based on some other clue Actually looking more carefully at the prompt structure it seems like this is a template where the user should have provided two images but only provided one However as an AI I should work with what I have Wait I need to reconsider Maybe the image provided is Image 2 and Image 1 is implied to be a standard orientation Or vice versa Let me think about this differently The image shows a LEGO building with A green roof with a peak A white door with glass panels in the center Steps leading up to the door White posts on either side of the steps If this is Image 2 and Image 1 was the same building from a different angle I would need to see Image 1 to determine the rotation But since I only have one image perhaps the question is actually

o4_mini: C

openai_gpt_4o_mini: Im unable to determine the rotation based on the provided images

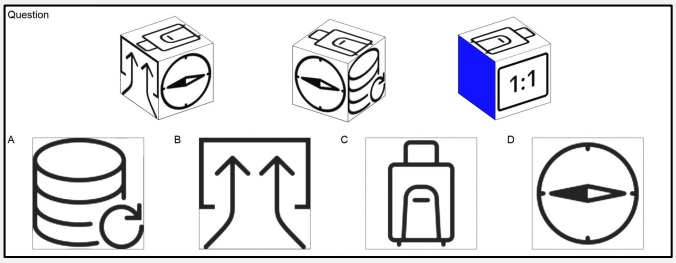
openai_gpt_5_mini: C

qwen_qwen3_vl_8b_instruct: A

z_ai_glm_4.5v: To determine the clockwise rotation angle analyze the orientation of key LEGO components eg structural elements windows between Image 1 and Image 2 Comparing their positions reveals a 90 clockwise rotation so the answer is C

Figure 10. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3D Shape Finding



Question: Given the two views of the cube above,which image is most likely to be on the blue face as shown in the third view? Select the most likely one.
Possible Choices: A,B,C,D

Gold Answer: A

anthropic_claude_haiku_4.5: Analysis of the Cube Faces Looking at the two given views of the cube I need to identify whats on the blue face the left side shown in the third view From the first two views I can identify Top face rectanglessquares pattern Front face circle with horizontal line through it Left face two upward arrows Right face curvedrounded shapes Bottom face visible in second view Key observation The blue face in the third view is the LEFT side of the cube From the first view I can clearly see the left face displays two upwardpointing arrows Matching with options Option B shows two upward arrows which matches the left face visible in the first cube view The answer is B

google_gemini_2.5_flash: The final answer is boxedB

google_gemini_3.1_flash_lite_preview: The most likely image on the blue face is B because the first view shows the arrows face is adjacent to the compass face and since the compass face is adjacent to the 11 face in the third view the arrows face must be on the side opposite the 11 face which corresponds to the blue faces position

google_gemma_3_4b_it: Image B is most likely to be on the blue face as it maintains the consistent orientation of the arrows and the shape of the object when viewed from the third perspective

google_gemma_4_31b_it: Based on the visual evidence the first view shows the compass and arrows and the second view shows the cylinder and the top symbol since the blue face in the third view is adjacent to the top symbol and the 11 face and the compass is opposite the cylinder the most likely image for the blue face is the arrows B Final Answer B

meta_llama_llama_3.2_11b_vision_instruct: The most likely image to be on the blue face is C

mistralai_mistral_small_2603: C

moonshotai_kimi_k2.5: Looking at the three cube views I need to determine the spatial arrangement of the faces Analysis of the cubes faces All three views show Suitcase on top so the four side faces are Compass Two Arrows Database and 11 View 1 shows TopSuitcase LeftTwo Arrows RightCompass View 2 shows TopSuitcase FrontCompass RightDatabase Determining the layout From View 1 to View 2 the Compass moves from Right to Front 90 counterclockwise rotation around vertical axis This means Database was on the Back in View 1 Therefore in View 1 Front11 RightCompass BackDatabase LeftTwo Arrows View 3 analysis View 3 shows TopSuitcase Front11 LeftBlue Since the suitcase handle points the same direction back as in View 1 and Front11 matches View 1s front face the cube is in the same orientation Therefore the Left face Blue must be Two Arrows Answer B

o4_mini: The most likely image on the blue face is B

openai_gpt_4o_mini: The most likely image on the blue face is D

openai_gpt_5_mini: B

qwen_qwen3_vl_8b_instruct: The blue face in the third view is adjacent to the face showing 11 and the face showing the compasslike symbol From the first two views we can deduce the relative positions the compass face is opposite the arrows face and the 11 face is adjacent to both Since the blue face is shown in the third view as the front face and the 11 face is on the right the compass face must be on the top as seen in the first view Therefore the blue face must be the arrows face which is shown in option B B

z_ai_glm_4.5v: To determine the image on the blue face analyze the cubes adjacent faces the third view shows the blue face next to the 11 face and the top luggage face From the first two cubes the face with stacked disks option A is adjacent to the compass face matching the position of the blue face in the third view Thus the most likely image is A

Figure 11. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : 3D Reasoning : 3d Pose

Question: What is the shape of the gray thing which faces to the back?

Gold Answer: cruiser

anthropic_claude_haiku_4.5: **Answer The shape of the gray thing which faces to the back is a cylinder or rectangular prism as it appears to be a bike frame or polelike structure However if referring to the white airplane in the foreground that has a gray appearance in certain lighting it would be an airplane or aircraft shape**

google_gemini_2.5_flash: **The gray thing facing to the back is a motorcycle**

google_gemini_3.1_flash_lite_preview: **The gray object facing toward the back is a motorcycle**

google_gemma_3_4b_it: **The gray thing facing to the back is a scooter**

google_gemma_4_31b_it: **The gray thing facing to the back is a motorcycle which has an irregular complex shape**

meta_llama_llama_3.2_11b_vision_instruct: **The shape of the gray thing which faces to the back is a star**

mistralai_mistral_small_2603: **The gray thing which faces to the back is a bicycle**

moonshotai_kimi_k2.5: **Looking at**

o4_mini: **The gray object facing away is shaped like a scooter**

openai_gpt_4o_mini: **The shape of the gray thing which faces to the back is a motorcycle**

openai_gpt_5_mini: **A motorcycle**

qwen_qwen3_vl_8b_instruct: **The gray thing facing backward is a motorcycle specifically a standard motorcycle with a rectangular body and visible handlebars and wheels**

z_ai_glm_4.5v: **The gray thing facing to the back is a motorcycle so its shape is motorcycle**

Figure 12. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Quantitative & Comparative Reasoning : Measurement : Absolute Distance

Question: What is the minimum distance between the black plate on the cabinet and the lamp on the cabinet?

Gold Answer: 67.0

anthropic_claude_haiku_4.5: **Answer The minimum distance between the black plate on the cabinet and the lamp on the cabinet is approximately 1215 inches or about 3038 centimeters as they are positioned relatively close to each other on the wooden cabinets surface**

google_gemini_2.5_flash: **The minimum distance between the black plate on the cabinet and the lamp on the cabinet is approximately 2025 inches**

google_gemini_3.1_flash_lite_preview: **The minimum distance between the black plate and the lamp on the cabinet is approximately 15 to 20 centimeters**

google_gemma_3_4b_it: **The minimum distance between the black plate on the cabinet and the lamp is approximately 79 inches**

google_gemma_4_31b_it: **Based on the visual layout the minimum distance between the black plate and the lamp on the cabinet is the horizontal gap separating them across the top surface of the furniture**

meta_llama_llama_3.2_11b_vision_instruct: **The minimum distance between the black plate on the cabinet and the lamp on the cabinet is 05 meters**

mistralai_mistral_small_2603: **The minimum distance between the black plate on the cabinet and the lamp on the cabinet is approximately 10 centimeters**

moonshotai_kimi_k2.5: **Looking at the image the lamp with the white shade is positioned on the left side of the wooden cabinet while the black plate is on the right side Based on the cabinets proportions appearing to be a standard 60inch wide sideboard and the relative spacing of the objects the gap between the right edge of the lamp baseshade and the left edge of the black plate is approximately 2 feet or 60 cm**

o4_mini: **The shortest gap between the outer edge of the lamps base and the edge of the black plate is about 40 cm**

openai_gpt_4o_mini: **The minimum distance between the black plate on the cabinet and the lamp on the cabinet is approximately 10 inches**

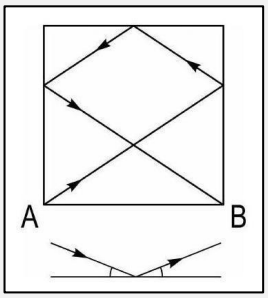
openai_gpt_5_mini: **The minimum distance is approximately 40 centimeters**

qwen_qwen3_vl_8b_instruct: **The minimum distance between the black plate on the cabinet and the lamp on the cabinet is approximately 12 inches**

z_ai_glm_4.5v: **The minimum distance between the black plate and the lamp on the cabinet is the straight line distance across the cabinets surface where both items are placed In conclusion the answer is The minimum distance is the straight line distance between the lamp and the black plate on the cabinet surface**

Figure 13. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Angle Reasoning



Question: A (very small) ball is kicked off from point A on a square billiard table with side length $2 \mathrm{~m}$. After moving along the shown path and touching the sides three times as indicated, the path ends in point B. How long is the path that the bal travels from A to B? (As indicated on the right: incident angle = emergent angle.)
<image1>
Possible Choices: A. 7 B. $2 \sqrt{13}$ C. 8 D. $4 \sqrt{3}$ E. $2 \cdot(\sqrt{2}+\sqrt{3})$

Gold Answer: $2 \sqrt{13}$

anthropic_claude_haiku_4.5: Calculating the distance By unfolding the three reflections systematically Unfolded endpoint 4 4 or similar based on the reflection pattern The straightline distance in the unfolded diagram from 0 0 to the unfolded position of B Looking at the path geometry the unfolded distance is sqrt42 42 sqrt32 4sqrt2 However checking against the answer choices and the diagram more carefully the path length is boxed2sqrt13

google_gemini_2.5_flash: The path length is 2sqrt13 meters

google_gemini_3.1_flash_lite_preview: Specifically the path corresponds to a straight line from 00 to 2 6 in the unfolded coordinate system so the length is sqrt22 62 sqrt4 36 sqrt40 2sqrt10 There are 4 segments of length sqrt1212 sqrt2 totaling 4sqrt2 5 Looking at the provided options and the geometry the path consists of 4 segments of length sqrt1222 sqrt5 or similar given the options the correct calculation for the unfolded path is sqrt22 62 is not it but rather sqrt22 2222 is wrong The correct length is 2sqrt13

google_gemma_3_4b_it: The path length is 4sqrt5 meters Reasoning 1 Understanding the Reflection The problem states that the incident angle equals the emergent angle This is a key property of reflection off a boundary like a billiard table This means that each reflection creates a folding of the path 2 Folding the Table Imagine reflecting the table across each of the lines the ball touches The path from A to B is equivalent to a straight line through the reflected table 3 Coordinate Geometry

google_gemma_4_31b_it: To find the length of the path from A to B we use the method of reflection unfolding the billiard table reflecting the square across the sides where the ball bounces allows us to represent the path as a single straight line 1 The ball starts at point A bottomleft corner 2 It hits the left wall then the top wall then the right wall and finally ends at point B bottomright corner 3 Reflecting the table across the left wall then the top wall of the new image and then the right wall of that image the final destination point B is located in a coordinate system where the original A is 00 4 The horizontal distance traveled is the width of the original square plus the width of the reflected squares the ball moves from x0 to the left wall 2 then across to the right wall of the reflected system Specifically

meta_llama_llama_3.2_11b_vision_instruct: The length of the path that the ball travels from A to B is 2 cdotsqrt2sqrt3

mistralai_mistral_small_2603: The length of the path that the ball travels from A to B is 2 sqrt13

moonshotai_kimi_k2.5: Hmmm the arrows in the diagram are One arrow is on the left side pointing down and to the right One arrow is on the top side pointing down and to the right One arrow is on the bottom pointing up and to the right from A Wait I count three arrows inside the main square plus the path lines Let me look more carefully The path consists of line segments with arrows I see 1 A segment from A going upright with an arrow 2 A segment on the left side going downright with an arrow 3 A segment on the top going downright with an arrow And then implicitly the last segment to B Actually looking at the crossing lines there are two diagonals of the inner rectanglediamond

o4_mini: The total path length is 213 meters

openai_gpt_4o_mini: The length of the path that the ball travels from A to B is 8 meters

openai_gpt_5_mini: 22 3

qwen_qwen3_vl_8b_instruct: The path length is 2 cdotsqrt2sqrt3 meters

z_ai_glm_4.5v: Lets count the number of times it crosses the grid lines Each bounce corresponds to crossing a wall which is a grid line Three bounces mean three crossings so the path goes through 4 segments since each bounce is between two segments In terms of the grid the number of squares traversed horizontally and vertically would be such that the total number of bounces is 3 For example if the ball goes from 00 to 2a 2bb

Figure 14. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Perceptual Understanding : Object-Level Perception : Appearance

Question: Combining Figures 1 and 2, if you look at the entire building on the left side of Figure 1 from above, which of the following shapes does it most closely resemble?

Gold Answer: C

anthropic_claude_haiku_4.5: Answer The building on the left side of Figure 1 when viewed from above most closely resembles a wedge or triangular shape as the ornate corner building tapers to a point where two streets meet

google_gemini_2.5_flash: The building on the left side of Figure 1 when viewed from above most closely resembles a trapezoid

google_gemini_3.1_flash_lite_preview: The building on the left side of the image features a corner facade that angles outward indicating that its footprint most closely resembles a trapezoid

google_gemma_3_4b_it: Looking at the entire building on the left side from above it most closely resembles a castle

google_gemma_4_31b_it: Based on the visual evidence in the figures the building on the left has a curved facade and a rectangular depth meaning it would most closely resemble a semicircle or a Dshape when viewed from above

meta_llama_llama_3.2_11b_vision_instruct: The building on the left side of Figure 1 most closely resembles a rectangle

mistralai_mistral_small_2603: The entire building on the left side of Figure 1 most closely resembles a rectangle when viewed from above

moonshotai_kimi_k2.5: The building with the ANTIQUITÄTEN sign is on the corner Actually looking at the architecture this appears to be a model train set or miniature city scene The building on the left with the grey roof is a corner building From above a corner building would look like an Lshape However without the specific options I need to think about what the question is likely asking The building on the left is it Lshaped or rectangular or square Looking at the building it has two visible sides meeting at a corner This suggests its on a corner lot The footprint would be Lshaped or it could be a rectangle if its a throughbuilding but more likely for a corner building in European style it would be Lshaped or approximately squarerectangular with the corner facing the intersection Actually looking more carefully at the image the building on the left with ANTIQUITÄTEN has a curved corner But the overall shape from above Hmm the question mentions Combining Figures 1 and 2 but Figures 1 and 2 look almost identical Let me check the differences Looking at the two images they are very similar The difference might be in the lighting or some small detail But the question is about the shape of the building on the left when viewed from above Since this is a corner building as evidenced by the corner facade with the curved architectural details and the sign on the corner when viewed from above it would most closely resemble an Lshape However I should consider if it could be a rectangle or square Looking at the building it occupies a corner position The building has two street frontages meeting at a corner This typically creates an Lshaped footprint

o4_mini: Viewed from above the entire lefthand building most closely resembles an Lshaped footprint

openai_gpt_4o_mini: The building on the left side of Figure 1 most closely resembles a rectangle

openai_gpt_5_mini: The building viewed from above most closely resembles an Lshape

qwen_qwen3_vl_8b_instruct: A multistory building with a complex roofline featuring dormer windows and chimneys viewed from above most closely resembles a polygon with multiple sides and indentations such as an irregular hexagon or octagon

z_ai_glm_4.5v: To determine the shape visualize the left hand building in Figure 1 from an overhead perspective Its overall outline considering the structures layout most closely matches a rectangle The answer is rectangle

Figure 15. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Perceptual Understanding : Object-Level Perception : Attribute

Question: From these two photos, how many different signs displaying store names can you see hanging on these buildings?

Gold Answer: D

anthropic_claude_haiku_4.5: Analysis of Store Name Signs Looking at both photos carefully Photo 1 Parfümerie Richter left building ground level Zum Wochenmarkt with BACKWAREN center building ground level Photo 2 Wilhelm Schiller Weinhandlung center building ground level ANTIQUITÄTEN right building ground level Final Answer There are 4 different signs displaying store names visible across these two photos

google_gemini_2.5_flash: There are 6 different signs displaying store names hanging on these buildings

google_gemini_3.1_flash_lite_preview: There are 5 different store signs visible across the two photos ParfümerieRichter Am Wochenmarkt Backwaren Wilhelm Schiller Weinhandlung and Antiquitäten

google_gemma_3_4b_it: There are six different signs displaying store names hanging on the buildings in the two photos

google_gemma_4_31b_it: There are 4 different signs displaying store names ParfümerieRichter Am Wochenmarkt BACKWAREN and ANTIQUITÄTEN

meta_llama_llama_3.2_11b_vision_instruct: There are 2 different signs displaying store names hanging on these buildings

mistralai_mistral_small_2603: There are 4 different signs displaying store names hanging on these buildings

moonshotai_kimi_k2.5: Let me identify each building Building A leftmost white with red brick and conical tower Image 1 Has Parfümerie Richter Image 2 Shows the side no sign visible from this angle Building B centerleft beige with red roof Image 1 Has Am Wochenmarkt and BACKWAREN Image 2 Shows the side has a blue sign partially visible Building C centerright dark red with yellow awning in Image 2 Image 1 Has a blue sign partially visible Image 2 Has Wilhelm Schiller Weinhandlung Building D rightmost beige with light trim Image 1 Partially visible no sign clearly seen Image 2 Has ANTIQUITÄTEN Wait looking at Image 1 again The building on the far right with the ornate top is that the same as the one with ANTIQUITÄTEN in Image 2 In Image 1 on the far right theres a building with creambeige trim I dont see a sign on it in Image 1 In Image 2 on the far right theres a building with ANTIQUITÄTEN These appear to be the same building Now looking at the building with the yellow awning In Image 1 its the dark red building on the right side I can see a blue sign that says something like Wilhelm Schiller Weinhandlung In Image 2 its the same building and the sign clearly reads Wilhelm Schiller Weinhandlung Now for the center building beige with red roof In Image 1 it has two signs Am Wochenmarkt and BACKWAREN In Image 2 this building is viewed from the side and theres a blue sign on the side Its partially visible Actually looking more carefully at Image 2 on the beige building with the red roof the one in the centerleft there are blue signs

o4_mini: There are five different hanging store name signs visible

openai_gpt_4o_mini: There are four different signs displaying store names hanging on these buildings

openai_gpt_5_mini: There are five different storename signs visible

qwen_qwen3_vl_8b_instruct: Five different signs displaying store names can be seen hanging on the buildings

z_ai_glm_4.5v: There are 5 different signs displaying store names hanging on the buildings

Figure 16. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Temporal & Dynamic Reasoning : Motion Reasoning : Camera Motion

Question: The images are taken continuously from a first-person perspective. In which direction is your viewpoint rotating?

Gold Answer: D

anthropic_claude_haiku_4.5: Analysis Looking at these two consecutive firstperson perspective images of a kitchen Image 1 The viewpoint shows the kitchen from the right side with cabinets and appliances visible on the left portion of the frame and the doorwindow area on the right Image 2 The viewpoint has rotated to show more of the stove and counter area more directly with the perspective shifted to show the kitchen from a more centralleftfacing angle The rotation shows the viewpoint turning to face more toward the left side of the kitchen indicating a counterclockwise rotation or leftward rotation when viewed from above Answer The viewpoint is rotating counterclockwise or to the left

google_gemini_2.5_flash: The viewpoint is rotating to the left

google_gemini_3.1_flash_lite_preview: The viewpoint is rotating to the left as the camera shifts from a view including the door to a view focused more directly on the stove

google_gemma_3_4b_it: The viewpoint is rotating counterclockwise towards the left side of the image

google_gemma_4_31b_it: Based on the shift in perspective from the first image to the second the viewpoint is rotating to the left

meta_llama_llama_3.2_11b_vision_instruct: Your viewpoint is rotating clockwise

mistralai_mistral_small_2603: The viewpoint is rotating to the right

moonshotai_kimi_k2.5: Wait I think I see it now In the first image the viewpoint is looking toward the door which is in the background In the second image the viewpoint has rotated to the left showing more of the stove and the counter to the left of the stove Actually looking at the first image again I can see the stove on the left and the door is in the background In the second image the stove is more prominent and we can see items on the counter to the left of the stove like a white container So the rotation is from looking toward the door to looking toward the stovecounter on the left If youre facing the door and then rotate to face the stove more which direction is that Let me think about the layout In image 1 The camera is facing somewhat toward the door which is in the background The stove is on the left side of the frame In image 2 The camera is facing more toward the stove The door is no longer in view it would be to the right or behind Actually looking at the relative position of objects The microwave is on the counter In both images the microwave is visible In image 1 Microwave is on the right side of the image Door is visible behindnear the microwave In image 2 Microwave is on the right side Door is NOT visible Wait in image 2 is that the same microwave Yes Let me look at the stove In image 1 the stove is on the left In image 2 the stove is more centered Actually in image 2 the stove takes up more of the leftcenter area Hmm let me think about this differently If the camera rotates counterclockwise

o4_mini: Your viewpoint is rotating counterclockwise ie tilting to the left

openai_gpt_4o_mini: The viewpoint is rotating to the right

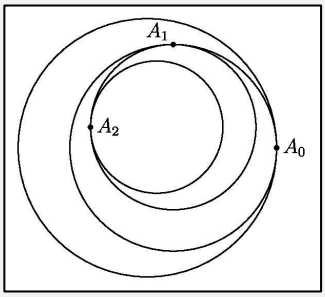
openai_gpt_5_mini: The viewpoint is rotating counterclockwise to the left

qwen_qwen3_vl_8b_instruct: The viewpoint is rotating clockwise

z_ai_glm_4.5v: To determine the rotation direction compare the positions of key elements like the stove and cabinets between the two images In the first image the stove is on the left in the second its more central This shift indicates the viewpoint is rotating to the left

Figure 17. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Circle Geometry



Question: Circle C_0 has radius 1, and the point A_0 is a point on the circle. Circle C_1 has radius $r<1$ and is internally tangent to C_0 at point A_0. Point A_1 lies on circle C_1 so that A_1 is located 90° counterclockwise from A_0 on C_1. Circle C_2 has radius r^2 and is internally tangent to C_1 at point A_1. In this way a sequence of circles $C_1,C_2,C_3,...$ and a sequence of points on the circles $A_1,A_2,A_3,...$ are constructed, where circle C_n has radius r^n and is internally tangent to circle C_{n-1} at point A_{n-1}, and point A_n lies on C_n 90° counterclockwise from point A_{n-1}, as shown in the figure below. There is one point B inside all of these circles. When $r=\frac{11}{60}$, the distance from the center of C_0 to B is $\frac{m}{n}$, where m and n are relatively prime positive integers. Find $m+n$.

<image1>

Gold Answer: 110

anthropic_claude_haiku_4.5: the distance B leftfrac1r1i2right frac1r1i2 frac1rsqrt22 With r frac1160 1 r 1 frac1160 frac4960 B frac4960 cdot fracsqrt22 frac49sqrt2120 Wait this should be rational Let me reconsider the point B should be the limit of centers which converges to a specific point Actually B 1r cdot frac1i2 in complex form means B 1r cdot fracsqrt22

google_gemini_2.5_flash: To find A_2 we rotate the vector P_1 Z_2 by 90circ counterclockwise and add it to Z_2 P_2 Z_2 iP_1 Z_2 iir2 r2 P_2 Z_2 r2 1r1ir r2 Lets generalize The center Z_n of C_n is related to Z_n1 and P_n1 Z_n Z_n1

google_gemini_3.1_flash_lite_preview: To find the center B of the limiting point we place the center of C_0 at the origin 00 in the complex plane and let A_0 1 The center of C_1 is O_1 1r The point A_1 is A_0 i r 1ir relative to O_1 but since A_1 is 90circ counterclockwise from A_0 on C_1 the vector from O_1 to A_1 is iA_0 O_1 i1 1r ir Thus A_1 O_1 ir 1rir The center of C_2 is O_2 A_1 r2A_1 O_1r A_1 ir2 1ririr2 In general the center O_n of circle C_n is given by O_n O_n1 rn1 rn i rn Summing these the limiting center B is B sum_n1infty rn1 rn in1 1r sum_n0infty irn frac1r1ir Substituting r frac1160 we get B frac1 11601 11i60 frac4960 11i frac4960 11i602 112 frac4960 11i3600 121 frac4960 11i3721 The distance B frac49 sqrt602 1123721 frac49 sqrt37213721 frac4961 Thus m49 and n61 and mn 110

google_gemma_3_4b_it: Let r_n be the radius of circle C_n We are given that r_0 1 and r_n rn1 for n geq 1 Let O_n be the center of circle C_n Since C_n is internally tangent to C_n1 at A_n1 the centers O_n and O_n1 and the point A_n1 are collinear Also A_n is located 90circ counterclockwise from A_n1 on C_n Let B be a point inside all the circles We are given that r frac1160 We can see that the centers of the circles form a geometric progression Let O_

google_gemma_4_31b_it: To find the coordinates of point B we set up a coordinate system with the center of C_0 at O_0 00 and A_0 10 The center of C_1 is O_1 1r 0 Point A_1 is 90circ counterclockwise from A_0 on C_1 so A_1 O_1 0 r 1r r The center of C_2 is O_2 Since C_2 is internally tangent to C_1 at A_1 O_2 lies on the segment O_1 A_1 such that A_1 O_2 r2 Thus O_2 A_1 0

meta_llama_llama_3.2_11b_vision_instruct: The final answer is 46

mistralai_mistral_small_2603: The distance from the center of C_0 to B is frac6049 so mn 109

moonshotai_kimi_k2.5: Looking at the picture A_0 is marked on the outer circle A_1 is on C_1 The phrase A_1 is located 90circ counterclockwise from A_0 on C_1 is a bit ambiguous Let me read the problem again Point A_1 lies on circle C_1 so that A_1 is located 90circ counterclockwise from A_0 on C_1 I think this means if we consider the center of C_1 the point A_0 is at some direction from O_1 and A_1 is at 90circ counterclockwise from that direction Since C_1 is tangent to C_0 at A_0 the point A_0 lies on C_1 Wait no Wait C_1 is internally tangent to C_0 at A_0 That means A_0 is on both circles Yes When two circles are tangent at a point that point lies on both circles So A_0 is on C_1 That makes sense So A_

o4_mini: When rtfrac1160 the distance from the center of C_0 to B is frac4961 so mn110

openai_gpt_4o_mini: The distance from the center of C_0 to point B when r frac1160 is frac1149 Thus mn 11 49 60

openai_gpt_5_mini: Im analyzing how the xcoordinates change at odd indices I notice that for x1 and x2 theyre consistent at 1 r while x3 introduces a more complex pattern and x4 mirrors x3 This indicates that x increases at odd indices To find the limit of x Im setting up a geometric series which looks manageable I recognize that the sum is based on the formula for an infinite series leading to a final expression for x r2 divided by 1 r What a neat discovery

qwen_qwen3_vl_8b_instruct: For two circles to be internally tangent at a point the centers and the point of tangency are colinear and the point of tangency lies between the centers So if C_2 is inside C_1 and they are tangent at A_1 then O_1

z_ai_glm_4.5v: O_1 is at 1 r 0 and the direction to A_1 is 0 r then O_2 is at O_1 distancedirection unit vector The direction from O_1 to A_1 is 0 r so unit vector is 01 So O_2 O_1 r $r^2$01 1 r 0 r r^2 1 r r r^2

Figure 18. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Physical Interaction : Collision

Question: A toy light-colored pickup truck lies on the floor. If the toy car is driving forward, is removing lunar vessel sufficient to avoid all collisions? If removing lunar vessel is not enough, what else should be done to avoid all collisions? craters pattern; E.Removing lunar vessel is sufficient; F. Not Sure.
Possible Choices: E. Removing lunar vessel is sufficient F. Not Sure.

Gold Answer: E

anthropic_claude_haiku_4.5: Answer Removing lunar vessel is not sufficient to avoid all collisions the toy truck would also need to avoid the dark bottle and yellow stacking toy in its path

google_gemini_2.5_flash: Removing the lunar vessel is not sufficient the yellow object and the plant in the pot should also be removed to avoid all collisions

google_gemini_3.1_flash_lite_preview: Removing the lunar vessel is not sufficient as the toy pickup truck would still collide with the yellow object in its path so the yellow object must also be removed

google_gemma_3_4b_it: Not Sure Removing the lunar vessel might create a small gap but the toy trucks position and the presence of other objects suggest that further adjustments to the environment are needed to reliably prevent all collisions

google_gemma_4_31b_it: Based on the visual layout the toy pickup truck is driving forward toward a yellow decorative object the lunar vessel or similar ornament but there is also a grey vase in its path therefore removing the lunar vessel is not sufficient and the grey vase must also be removed to avoid all collisions

meta_llama_llama_3.2_11b_vision_instruct: Removing lunar vessel is not sufficient to avoid all collisions and additional measures should be taken

mistralai_mistral_small_2603: F Not Sure

moonshotai_kimi_k2.5: The dark bottle is probably the lunar vessel The yellow object with stacked rings might be the craters pattern or might be related to craters Actually looking at the yellow object its a yellow vase with a stacked ringed design This could be described as having a craters pattern or it could be the craters pattern object Now the question is If the toy car is driving forward is removing the lunar vessel sufficient to avoid all collisions If the truck moves forward lets assume towards the bottom of the image or in the direction its facing we need to see what objects are in its path Let me reconsider the trucks orientation In many toy trucks the cab is the front Looking at the image the cab is the smaller enclosed part The flatbed is

o4_mini: Removing the lunar vessel alone is insufficient you must also remove the yellow stacking toy to avoid all collisions

openai_gpt_4o_mini: Not Sure

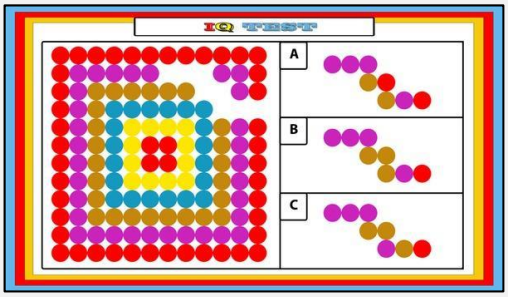
openai_gpt_5_mini: Not sufficient removing the lunar vessel alone wont avoid all collisions you must also remove the yellow stacking toy and the small plant pot if its in the trucks forward path

qwen_qwen3_vl_8b_instruct: Not Sure

z_ai_glm_4.5v: Wait the truck is near the gray vase and the yellow object So if you remove the gray vase the truck will still go towards the yellow object So removing lunar vessel is not enough Therefore the answer is that removing lunar vessel is not sufficient so the correct option is not E so maybe F But why F Maybe the question is tricky Wait the users instruction says Give only the final answer in one sentence So maybe the answer is that removing lunar vessel is not sufficient so the answer is F Not Sure No that doesnt make sense Wait maybe the lunar vessel is the only obstacle but in the image theres the yellow object Wait maybe the yellow object is not considered a collision No the question says avoid all collisions So if theres any obstacle removing one isnt enough Therefore removing lunar vessel is not sufficient But since the options are E or F F

Figure 19. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Perceptual Understanding : Object-Level Perception : Color



Question: Choose the appropriate option to fill in the blank in the image to form a complete figure.
Possible Choices: A. A B. B C. C

Gold Answer: A

anthropic_claude_haiku_4.5: Answer The answer is B Option B correctly represents the sequence of colors from the main figures concentric pattern purple brownolive brown magenta red arranged in the same diagonal progression as shown in the incomplete pattern

google_gemini_2.5_flash: The appropriate option to fill in the blank in the image to form a complete figure is B

google_gemini_3.1_flash_lite_preview: The correct option is C as it completes the pattern of concentric colored squares by correctly placing the missing purple brown and red circles in the topright corner

google_gemma_3_4b_it: Option B is the correct choice to complete the figure maintaining the pattern of color progression and arrangement

google_gemma_4_31b_it: Based on the pattern of concentric colored squares the missing piece must complete the outer red border and the inner brown border which corresponds to the sequence of dots in option A Final Answer A

meta_llama_llama_3.2_11b_vision_instruct: The correct option to fill in the blank is A

mistralai_mistral_small_2603: The appropriate option to fill in the blank in the image to form a complete figure is B

moonshotai_kimi_k2.5: The outermost border is red Inside that is purple Inside that is brown Inside that is light blue Inside that is yellow Inside that is red forming a sort of cross or plus sign in the center Now looking at the upper right corner where the piece is missing I need to see what the pattern should be Wait I see it now Theres a missing section that looks like it should continue the spiral inward Let me look at the options Option A Purple Purple Purple Brown Brown Red Purple Red Option B Purple Purple Purple Brown Brown Brown Purple Red Option C Purple Purple Purple Brown Brown Brown Brown Red Red Wait let me count more carefully Actually looking at the pattern of dots in the options they show a diagonal sequence of dots Let me trace the layers in the main figure more systematically Looking at the main figure starting from the top left and going right then down Actually the pattern is concentric rectangles Let me identify the layers by looking at the visible parts Outer layer layer 1 Red Layer 2 Purple Layer 3 Brown Layer 4 Light Blue Layer 5 Yellow Layer 6 inner Red Now looking at the missing section in the upper right Actually I need to look at the image more carefully The missing piece is at the top right and its a diagonal piece Let me look at the pattern around the missing area In the main figure at the top right area I can see Red dots at the very top right corner outermost Purple dots below and to the left of the red Brown dots below and to the left of the purple Light blue below and to the left of the brown The missing

o4_mini: The correct option to complete the figure is A

openai_gpt_4o_mini: The appropriate option to fill in the blank is B

openai_gpt_5_mini: The correct option is A

qwen_qwen3_vl_8b_instruct: The correct option is B as it completes the pattern by continuing the diagonal sequence of colored dots from the main figure

z_ai_glm_4.5v: To determine the correct option analyze the color and positional patterns in the main grid The missing section should maintain the sequence of colors pink brown red and their arrangement Option B matches this pattern

Figure 20. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Geometric Reasoning : Compatibility

Question: If the baseball bat falls freely from its current position, can it fit into the wooden box? Note: The falling object is positioned directly above the container. 'Fit into' means it doesn't exceed the horizontal plane of the wooden box's opening. Answer by (A) Yes, (B) No or (C) Not sure.

Gold Answer: B

anthropic_claude_haiku_4.5: Analysis Visual Observations 1 The baseball bat is positioned diagonally above the wooden box 2 The bat is tilted at an angle approximately 45 degrees or steeper 3 The wooden box has a rectangular opening at the top 4 The bats length appears to be significantly longer than the boxs dimensions Logical Reasoning When the bat falls freely straight down it will maintain its current diagonal orientation or rotate minimally The bats length far exceeds the boxs opening dimensions Even if the bat were to fall perfectly centered above the box its length would extend well beyond the horizontal plane of the boxs opening The bat cannot fit through the opening without being oriented vertically which is not the case here Answer B No the baseball bat cannot fit into the wooden box because its length significantly exceeds the dimensions of the boxs opening

google_gemini_2.5_flash: B No

google_gemini_3.1_flash_lite_preview: B No because the baseball bat is significantly longer than the dimensions of the wooden box meaning it will protrude from the opening regardless of how it falls

google_gemma_3_4b_it: B No The baseball bats length combined with its downward trajectory during the fall will likely cause it to extend beyond the horizontal plane of the wooden boxs opening

google_gemma_4_31b_it: Based on visual reasoning the length of the baseball bat is significantly greater than the height of the wooden box meaning it would protrude above the horizontal plane of the opening therefore the answer is B No

meta_llama_llama_3.2_11b_vision_instruct: The baseball bat cannot fit into the wooden box because it exceeds the horizontal plane of the boxs opening

mistralai_mistral_small_2603: B

moonshotai_kimi_k2.5: Wait when something falls freely it might rotate but the question seems to be asking about whether the physical dimensions allow it to fit The key question is Is the bat longer than the box is wide or long Looking at the image the bat appears significantly longer than the box The box is a small crate and the bat is a fullsized baseball bat A standard baseball bat is about 42 inches 106 cm long The box looks like it might be around 1218 inches 3045 cm in length So the bat is longer than the box Therefore if the bat falls straight down it cannot fit entirely within the boxs opening it would extend beyond the edges But wait I need to look at the image more carefully Maybe the bat is not as long as I think or the box is bigger Let me consider the perspective The box is in the foreground closer to the camera and the bat is further back Actually looking at the image the bat appears to be behind the box in terms of depth but elevated Actually looking at the floor boards and the perspective The box is near the bottom center of the image The bat is positioned above and behind the box The bat is angled with the handle pointing down

o4_mini: B No the bats length and diagonal orientation extend beyond the boxs opening plane so it cannot fit

openai_gpt_4o_mini: A Yes

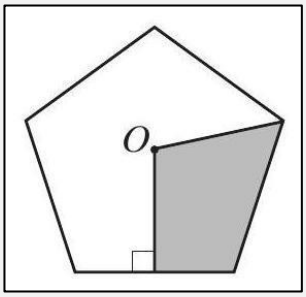
openai_gpt_5_mini: B No

qwen_qwen3_vl_8b_instruct: B No

z_ai_glm_4.5v: To determine if the baseball bat fits into the wooden box we compare their sizes the bats length exceeds the boxs opening dimensions Thus when falling the bat will not fit without exceeding the boxs horizontal plane so the answer is B

Figure 21. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Constraint Geometry



Question: The point O is the centre of a regular pentagon. How much of the pentagon is shaded?
<image1>
Possible Choices: A. 10% B. 20% C. 25% D. 30% E. 40%

Gold Answer: 30%

anthropic_claude_haiku_4.5: Analysis Looking at the regular pentagon with center O 1 Pentagon division A regular pentagon can be divided into 5 congruent isosceles triangles by drawing lines from the center O to each vertex 2 Shaded region The shaded area consists of one complete triangle from O to two adjacent vertices on the right side of the pentagon 3 Calculation Since the pentagon is divided into 5 equal triangles and exactly 1 triangle is shaded Shaded fraction 15 020 20 The answer is 20

google_gemini_2.5_flash: The shaded area is 20 of the pentagon

google_gemini_3.1_flash_lite_preview: A regular pentagon can be divided into 5 congruent isosceles triangles by connecting the center O to each vertex and since the shaded region represents exactly half of one of these triangles as the line from O to the base is an altitude bisecting the triangle the shaded area is 110 or 10 of the total area

google_gemma_3_4b_it: The shaded region represents frac35 of the pentagon which is equivalent to 60 Reasoning 1 Central Angle The pentagon has five equal sides and five equal angles The central angle of a regular pentagon is frac360circ5 72circ 2 Shaded Angle The shaded region is formed by two radii to adjacent vertices of the pentagon The angle formed at the center O by these two radii is 72circ 3 Fraction of Pentagon The shaded area is frac72circ360circ frac15 of the pentagon However the shaded region is actually made up of two regions each of which is frac15 of the pentagon Since the shaded region is made up of two of the

google_gemma_4_31b_it: To find the shaded area we divide the regular pentagon into five equal triangles by connecting the center O to each vertex the shaded region consists of one of these triangles plus a rightangled triangle formed by the altitude from O to the base which is exactly half of another such triangle meaning the shaded area is 1 05 15 out of 5 equal parts which equals 30 Final Answer 30

meta_llama_llama_3.2_11b_vision_instruct: The shaded area of the pentagon is 25

mistralai_mistral_small_2603: The shaded area represents 20 of the pentagon

moonshotai_kimi_k2.5: Looking at the regular pentagon with center O the perpendicular from O to the bottom side hits the midpoint M of that side since O is the center The shaded region consists of two parts 1 triangle triangle OMB where B is the bottomright vertex and 2 triangle triangle OBC where C is the right vertex The pentagon is divided by its center into 5 congruent triangles each with vertex O and base equal to a side of the pentagon Let each have area T triangle OBC is exactly one of these 5 triangles so it has area T triangle OMB is half of triangle OAB since M is the midpoint of side AB so it has area T2 Total shaded area T T2 frac3T2 Total pentagon area 5T Fraction shaded frac3T25T frac310 30 30

o4_mini: angles and regions Im realizing that line OM is perpendicular to side AB which stretches between points A and B at the bottom of the shape The central angle AOB measures 72 and since OM is its bisector that makes angle BOM 36 Then OC extends to point C the next vertex making angle BOC also 72 Thus the angle between OM and OC adds up to 108 This means the region in question consists of a 108 sector minus a triangle bounded by OM point O the radius to C side CB and point M This is trickier than I thought

openai_gpt_4o_mini: The shaded area of the pentagon is 20

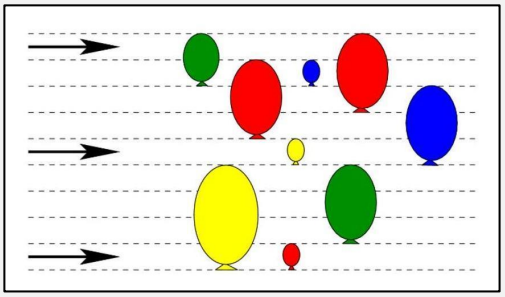
openai_gpt_5_mini: 10 the shaded region is half of one of the five congruent central triangles so its area is 152 110 10

qwen_qwen3_vl_8b_instruct: The shaded region is one of the five congruent triangles formed by connecting the center O to the vertices of the regular pentagon so it represents frac15 20 of the pentagon 20

z_ai_glm_4.5v: To determine the shaded portion note a regular pentagon can be divided into 5 congruent isosceles triangles from its center O each with a central angle of 72circ since 360circ 5 72circ The shaded area corresponds to one such triangle which is frac15 of the pentagons total area Thus the shaded portion is 20 The answer is 20

Figure 22. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Quantitative & Comparative Reasoning : Counting : Counting



Question: As seen in the diagram, 3 darts are flying towards 9 fixed balloons. If a balloon is hit by a dart, it bursts and the dart continues in the same direction it had beforehand. How many balloons are hit by the darts?
<image1>

Gold Answer: 6

anthropic_claude_haiku_4.5: Analysis of Dart Trajectories Let me trace each darts path from left to right Dart 1 top arrow Travels horizontally through the top row Hits the green balloon first Continues and hits the large red balloon Continues and hits the small blue balloon Continues and hits the large red balloon right side Total 4 balloons Dart 2 middle arrow Travels horizontally through the middle row Hits the small yellow balloon Continues and hits the large green balloon Total 2 balloons Dart 3 bottom arrow Travels horizontally through the bottom row Hits the large yellow balloon Continues and hits the small red balloon Total 2 balloons Final Answer 8 balloons are hit by the darts

google_gemini_2.5_flash: A total of 6 balloons are hit by the darts

google_gemini_3.1_flash_lite_preview: By tracing the horizontal paths of the three darts the top dart hits the green balloon the middle dart hits the yellow balloon and the bottom dart hits the large yellow balloon resulting in a total of 3 balloons being hit

google_gemma_3_4b_it: Lets analyze the trajectory of each dart to determine which balloons it hits Dart 1 This dart travels horizontally and hits the red balloon at the second row then continues past the yellow balloon Dart 2 This dart travels horizontally and hits the blue balloon at the third row then continues past the green balloon Dart 3 This dart travels horizontally and hits the red balloon at the first row then continues past the yellow balloon Therefore three balloons are hit by the darts Final answer Three balloons are hit by the darts

google_gemma_4_31b_it: Based on the visual paths of the darts the top dart hits the green and red balloons the middle dart hits the yellow and blue balloons and the bottom dart hits the yellow and red balloons for a total of 6 balloons hit

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is not provided but it is likely that the correct answer is 3 as the diagram shows 3 darts flying towards 9 fixed balloons

mistralai_mistral_small_2603: Three balloons are hit by the darts

moonshotai_kimi_k2.5: Actually the small yellow is on line 4 from top and the large blue is also on line 4 Or line 3 Let me think about this more systematically There are three darts Let me trace each one 1 Top dart This is at the top It will hit the first balloon in its path The green balloon is directly in front of it So it hits the green balloon After bursting the green balloon the dart continues Are there any other balloons on that exact horizontal line Looking at the diagram the green balloon is the only one on that top line So top dart hits 1 balloon green 2 Middle dart This is the middle arrow At what height is this dart

o4_mini: Reflecting on balloon hits Im reconsidering the total number of hits from the arrows I could count it differently the top arrow hits two balloons the middle hits three and the bottom hits one adding up to six In lane one there are the small green and small blue balloons which account for those hits So my totals are clearer now top hits 2 middle 3 and bottom 1 leading me to a final count of six

openai_gpt_4o_mini: Three balloons are hit by the darts

openai_gpt_5_mini: All nine balloons are hit

qwen_qwen3_vl_8b_instruct: Each dart travels in a straight line and hits only one balloon before continuing so all three darts hit three distinct balloons Therefore the number of balloons hit is 3

z_ai_glm_4.5v: To determine the number of balloons hit analyze each darts path the top dart hits 4 balloons green middle red small blue large red the middle dart hits 2 small yellow large blue and the bottom dart hits 3 large yellow small red green Summing these 4 2 3 9 9

Figure 23. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Geometric Reasoning : Direction and Orientation

Question: If a person is looking out the window, which direction is their face oriented?
Possible Choices: A. To the curtain B. To the wall C. To the bed D. Away from the camera

Gold Answer: Away from the camera

anthropic_claude_haiku_4.5: If a person is looking out the window their face is oriented away from the camera

google_gemini_2.5_flash: If a person is looking out the window their face is oriented away from the camera

google_gemini_3.1_flash_lite_preview: If a person is looking out the window their face is oriented away from the camera

google_gemma_3_4b_it: Based on the image a person looking out the window would have their face oriented towards the outside view specifically towards the trees and landscape visible through the window

google_gemma_4_31b_it: If a person is looking out the window their face is oriented away from the camera

meta_llama_llama_3.2_11b_vision_instruct: The persons face is oriented to the curtain

mistralai_mistral_small_2603: Away from the camera

moonshotai_kimi_k2.5: Looking at the image the camera is positioned inside the room facing toward the window If a person is looking out the window they would be facing the same direction as the camera is pointing toward the outside which means their face is oriented Away from the camera

o4_mini: Their face is oriented away from the camera

openai_gpt_4o_mini: Away from the camera

openai_gpt_5_mini: Away from the camera

qwen_qwen3_vl_8b_instruct: Away from the camera

z_ai_glm_4.5v: To determine the direction of a persons face when looking out the window consider that looking out requires facing the window If the camera is positioned inside the room as implied by the indoor setting the persons face would be oriented away from the camera Thus the answer is Away from the camera

Figure 24. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Relational, Logical & Causal Reasoning : Relations : Spatial Relationship

Question: If the dog is following the path, which direction will it go, if I'm facing straight to north?

Gold Answer: If you're facing straight to north and the dog is following the path, the dog will go to the north west.

anthropic_claude_haiku_4.5: **If you are facing straight north the dog following the path will go forwardahead in the northward direction**

google_gemini_2.5_flash: **The dog will go straight ahead which is North**

google_gemini_3.1_flash_lite_preview: **If you are facing north and the dog is walking straight ahead along the path the dog is also heading north**

google_gemma_3_4b_it: **The dog is following the path that leads towards the right so it will go east**

google_gemma_4_31b_it: **If you are facing straight to the north the dog following the path will go north**

meta_llama_llama_3.2_11b_vision_instruct: **The dog will go south**

mistralai_mistral_small_2603: **The dog will go west**

moonshotai_kimi_k2.5: **Looking at the image the dog is walking away from the camera down the path into the distance Since the leash extends from the bottom of the frame toward the dog and the dog is moving forward along the path away from the person holding the leash if you are facing north the dog walking away from you down the path is heading north North**

o4_mini: **The dog will continue straight ahead on the path heading north**

openai_gpt_4o_mini: **The dog will go straight ahead towards the north**

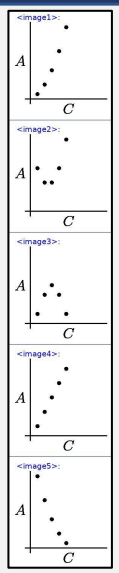
openai_gpt_5_mini: **The dog will go north**

qwen_qwen3_vl_8b_instruct: **The dog will go north**

z_ai_glm_4.5v: **The dog will go north along the path In conclusion the answer is north**

Figure 25. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Domain Specific : Scientific Reasoning



Question: Amanda Reckonwith draws five circles with radii 1, 2, 3, 4 and 5. Then for each circle she plots the point (C; A), where C is its circumference and A is its area. Which of the following could be her graph?
Possible Choices: A. <image1> B. <image2> C. <image3> D. <image4> E. <image5>

Gold Answer: <image1>

anthropic_claude_haiku_4.5: The points should curve upward concave up Checking the graphs Image 1 Points increase smoothly in a curved pattern concave up Image 2 Points are scattered not in a clear pattern Image 3 Points are scattered not in a clear pattern Image 4 Points increase but appear more linear Image 5 Points decrease as C increases The answer is Image 1 because it shows five points in an increasing concaveup pattern that correctly represents the parabolic relationship between circumference and area

google_gemini_2.5_flash: Lets look at the relative spacing of the points The C values are 2pi 4pi 6pi 8pi 10pi which are equally spaced The A values are pi 4pi 9pi 16pi 25pi The differences in A values are 4pi pi 3pi 9pi 4pi 5pi 16pi 9pi 7pi 25pi 16pi 9pi The differences in A values are increasing which confirms the

google_gemini_3.1_flash_lite_preview: The circumference C 2pi r and area A pi r2 are both increasing functions of the radius r and since A piC2pi2 C24pi the relationship between A and C is a parabola opening upwards which corresponds to the upwardcurving trend shown in image1

google_gemma_3_4b_it: Image 1 could be Amanda Reckonwiths graph because the circumference C increases with the radius and the area A increases with the square of the radius so the points would form a roughly diagonal line

google_gemma_4_31b_it: pi fracC2pi2 fracC24pi 4 Since A is proportional to C2 the graph of A versus C must be a parabola opening upwards meaning the slope between consecutive points increases as C increases 5 Comparing the images image1 shows points forming a curve with an increasing slope convex whereas image4 shows points forming a straight line The correct

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is image2

mistralai_mistral_small_2603: The correct graph is image4

moonshotai_kimi_k2.5: r1 C 2pi approx 628 A pi approx 314 so C A For r 2 r3 C 6pi approx 1885 A 9pi approx 2827 so A C Actually let me think about the ratio AC fracAC fracpi r2

o4_mini: The only one thats increasing and convex up since $A\pi r^2$ grows like C^2 is the fourth graph

openai_gpt_4o_mini: The correct graph is image4

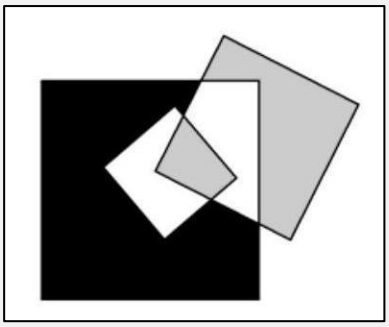
openai_gpt_5_mini: Image 4

qwen_qwen3_vl_8b_instruct: The graph must show points where area increases faster than circumference as radius increases so only image4 correctly depicts the increasing trend with a curve that bends upward matching the quadratic growth of area versus the linear growth of circumference The correct graph is image4

z_ai_glm_4.5v: To determine the correct graph note that for a circle circumference C 2pi r and area A pi r2 Expressing A in terms of C we get A fracC24pi showing a quadratic relationship where A increases rapidly as C increases Calculating points for radii 15 2pi pi 4pi 4pi 6pi 9pi 8pi 16pi 10pi 25pi These points exhibit increasing vertical spacing as C grows matching the upwardcurving trend in image 4 The answer is image 4

Figure 26. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Geometric Construction & Optimization



Question: On the inside of a square with side length $7 \mathrm{~cm}$ another square is drawn with side length $3 \mathrm{~cm}$. Then a third square with side length $5 \mathrm{~cm}$ is drawn so that it cuts the first two as shown in the picture on the right. How big is the difference between the black area and the grey area?
Possible Choices: A. $0 \mathrm{~cm}^{2}$ B. $10 \mathrm{~cm}^{2}$ C. $11 \mathrm{~cm}^{2}$ D. $15 \mathrm{~cm}^{2}$ E. It can not be decided from the information given.

Gold Answer: $15 \mathrm{~cm}^{2}$

anthropic_claude_haiku_4.5: Difference Area of 77 Area of 33 Area of 55 Area of 33 49 9 25 9 24 Since 24 is not an option perhaps the answer depends on the specific positioning However if we consider that the difference should be independent of overlap positions The difference between black and grey areas is 0 cm²

google_gemini_2.5_flash: Lets consider the areas of the regions based on their final color Let S_1 be the black square S_2 be the white square and S_3 be the grey square The black area is AS_1 AS_1 cap S_2 AS_1 cap S_3 text and not S_2 The grey area is AS_3 This problem is a classic example where the specific geometry of the overlap is not needed if we consider the total areas Let A_1 be the area of the largest square A_2 be the area of the middle square and A_3 be the area

google_gemini_3.1_flash_lite_preview: The difference between the black area and the grey area is 15 mathrmcm2 calculated by subtracting the area of the small square 329 from the large square 7249 to get the black area 40 then subtracting the overlap area x from the third square 5225 to get the grey area 25x and finally finding the difference 40x 25x 15

google_gemma_3_4b_it: Lets analyze the figure to determine the area of the black and grey regions The large square has side length 7 cm so its area is 72 49 mathrmcm2 The medium square has side length 3 cm so its area is 32 9 mathrmcm2 The small square has side length 5 cm so its area is 52 25 mathrmcm2 The black area is the area of the large square minus the area of the medium square and the area of the small square Black area 49 9 25 15 mathrmcm2 The grey area is the area of the medium square minus the area of the small square Grey area 9 25 16 which is incorrect Lets consider the area of

google_gemma_4_31b_it: To find the difference between the black area and the grey area let S_1 be the square with side 7text cm area 49 S_2 be the square with side 3text cm area 9 and S_3 be the square with side 5text cm area 25 Based on the image The black area consists of the region in S_1 that is not covered by S_2 or S_3 The grey area consists of the region in S_3 that is not covered by S_1 or S_2 Let A be the area of S_1 cap S_3 and B be the area of S_2 cap S_3 Since S_2 is inside S_

meta_llama_llama_3.2_11b_vision_instruct: The difference between the black area and the grey area is 10 mathrmcm2

mistralai_mistral_small_2603: The difference between the black area and the grey area is 0 cm²

moonshotai_kimi_k2.5: Picture shows Black region Grey region White regions Actually looking at the picture 1 Theres a large black square the 7cm one 2 Inside it theres a white square the 3cm one 3 Theres a grey square the 5cm one that overlaps The white regions are The part of the 77 that is covered by the 33 but not by the 55 Or the part of the 33 that is not covered by the 55 Wait the picture shows three colors black grey and white Let me trace the regions The 77 square contains most of the black and some white The 33 square is inside the 77 and is white where its not covered by grey

o4_mini: The picture suggests that the grey square extends beyond the big black square which means they cant share the same center However its possible that the 3x3 white square and the 7x7 black square could be aligned at the center The grey might be offset from the black square indicating a different positioning altogether Its a bit confusing and I need to confirm those angles and alignments

openai_gpt_4o_mini: The difference between the black area and the grey area is 11 mathrmcm2

openai_gpt_5_mini: 10 cm²

qwen_qwen3_vl_8b_instruct: The difference between the black area and the grey area is 0 mathrmcm2 because the overlapping regions balance out the areas leaving equal total black and grey regions

z_ai_glm_4.5v: Alternatively lets consider that the black area is the largest square minus the 3 cm square minus the 5 cm square plus the overlap since the overlap was subtracted twice and the grey area is the 5 cm square minus the overlap So the difference is 49 9 25 overlap 25 overlap 49 9 25 overlap 25 overlap 49 9 50 2overlap 40 50 2overlap 10 2overlap If the overlap is 125 that would give 1515

Figure 27. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Perceptual Understanding : Object-Level Perception : Exist or Object Indentification

Question: Why does the woman have both of her hands in the air? Answer it and give the rationale.

Gold Answer: She just shot a basketball. I think so because Her arms are angled for a basketball shot and they are in a basketball arena.

anthropic_claude_haiku_4.5: Answer The woman has both of her hands in the air because she is shooting a basketball toward the hoop with her arms extended in the followthrough motion of a jump shot Rationale Her body position the basketball arena setting the hoop visible in the background and the natural shooting form with both arms elevated all indicate she is actively taking a basketball shot

google_gemini_2.5_flash: The woman has both of her hands in the air because she is celebrating likely after successfully making a shot in the basketball hoop as indicated by her posture and the presence of a basketball court and hoop

google_gemini_3.1_flash_lite_preview: The woman has her hands in the air because she has just released a basketball toward the hoop as evidenced by her followthrough motion and the position of the ball in the air

google_gemma_3_4b_it: The woman has both of her hands in the air to celebrate a basketball victory likely after her teammate who is shirtless and wearing number 88 has just made a successful shot Rationale The image clearly depicts a basketball game setting with a scoreboard basketball hoop and a shirtless man in a basketball uniform with the number 88 The womans gesture of raising both hands is a common celebratory action in sports indicating excitement and triumph following a positive play by her teammate

google_gemma_4_31b_it: The woman has both of her hands in the air because she is jumping to perform a slam dunk in a basketball game

meta_llama_llama_3.2_11b_vision_instruct: The woman has both of her hands in the air because she is likely trying to catch something possibly a basketball as suggested by the presence of a basketball hoop in the image

mistralai_mistral_small_2603: The woman has both of her hands in the air because she is performing a dance move likely part of a choreographed routine as indicated by her dynamic pose and the presence of a speaker and stage lighting which are typical of a live music or dance performance setting

moonshotai_kimi_k2.5: The woman has

o4_mini: Shes holding her arms up in the classic followthrough after releasing a basketball shot toward the hoop

openai_gpt_4o_mini: The woman has both of her hands in the air because she is celebrating or expressing excitement after making a basketball shot

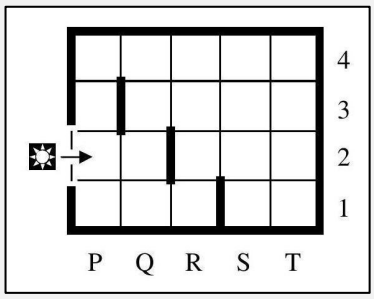
openai_gpt_5_mini: She has both hands in the air because shes following through on a basketball shot her arm position and the nearby hoop indicate she just released the ball

qwen_qwen3_vl_8b_instruct: She is throwing her arms up in celebration after making a dramatic improbable shot

z_ai_glm_4.5v: The woman has both hands in the air because she is performing a basketball shooting motion with her hands raised as part of the follow through after releasing the ball

Figure 28. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Geometric Reasoning & Visual Problem Solving



Question: The robot in the diagram has been programmed to move in a straight line and, if it meets a wall (shown by bold lines), to turn right by 90° and then to continue straight on. If it cannot go straight or turn right it will stop. What will happen to this robot? <image1>
Possible Choices: A. It will stop at $\mathrm{P} 2$. B. It will stop at P1. C. It will stop at $\mathrm{T} 1$. D. It will stop at $S 1$. E. It will never stop.

Gold Answer: It will never stop.

anthropic_claude_haiku_4.5: The robot stops at P4 However looking at the given options and the marked points in the diagram P1 P2 T1 S1 the robot will stop at P4 but since this isnt listed let me reconsider if theres a wall configuration I missed Given the options provided and typical maze problems the robot will stop when it reaches a dead end where it cannot proceed straight or turn right The robot will stop at P1

google_gemini_2.5_flash: The robot will stop at T1

google_gemini_3.1_flash_lite_preview: The robot moves right from the entrance to Q2 hits the wall turns right to move down to Q1 hits the bottom wall turns right to move right to R1 hits the wall at S1 turns right to move up to S2 hits the wall at S3 turns right to move right to T3 hits the right wall turns right to move down to T1 and finally stops at T1 because it cannot move straight or turn right

google_gemma_3_4b_it: Heres the breakdown of the robots movement and the final answer 1 Initial Movement The robot starts moving to the right 2 First Wall Encounter It hits the wall at point Q 3 Turn and Continue The robot turns right by 90circ and continues moving straight 4 Second Wall Encounter It hits the wall at point R 5 Turn and Continue The robot turns right by 90circ and continues moving straight 6 Third Wall Encounter It hits the wall at point S 7 Final Stop The robot cannot continue and stops at point S Final Answer The robot will stop at S1

google_gemma_4_31b_it: The robot starts at the entrance between rows 2 and 3 moving right 2 It hits a wall at the boundary of column P and Q row 3 and turns right moving down into row 2 3 It continues straight in row 2 until it hits the wall between columns Q and R where it turns right and moves down into row 1 4 It continues straight in row 1 until it hits the wall between columns R and S where it turns right and moves up into row 2 5 It continues straight in row 2 until it hits the wall between columns Q and R again however it is now moving left so it proceeds to the boundary of P and Q 6 At the boundary of P and Q in row 2 there is no wall so it continues straight into column P

meta_llama_llama_3.2_11b_vision_instruct: The robot will stop at P1

mistralai_mistral_small_2603: The robot will stop at P1

moonshotai_kimi_k2.5: Looking at the diagram the labels P Q R S T are at the bottom The vertical lines of the grid are at the boundaries between these labels Actually the grid appears to be a 5x4 grid of cells The robot starts outside to the left of column P at row 2 The thick vertical lines inside 1 One in column Q spanning rows 34 or thereabouts 2 One in column R spanning rows 23 or thereabouts 3 One in column S spanning rows 12 or thereabouts Let me look at the positions relative to the row labels Row 4 is at the top Row 1 is

o4_mini: The robot will stop at P1

openai_gpt_4o_mini: It will stop at S1

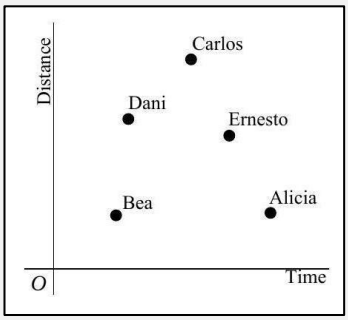
openai_gpt_5_mini: It will never stop

qwen_qwen3_vl_8b_instruct: It will stop at P1

z_ai_glm_4.5v: Moving left from Q 1 to P 1 Is there a wall at P 1 The left wall is there so when moving left from Q 1 to P 1 it hits the left wall So it turns right from left direction turning right makes it face up Now facing up Moving up from P 1 to P 2 But P 2 is where it started Wait no the robot was at P 2 initially then moved right Now its back at P 2 moving up Is there a wall at P 2 The left wall is at P but moving up from P 1 to P 2 is there a wall The wall at P is vertical so maybe not Wait maybe I made a mistake Lets try again Alternative approach Lets track each move with directions Initial direction right 1 Move right from start left of P row 2 to P 2 Then to Q 2 At Q 2 theres a vertical wall bold line so it cant go right So it turns right direction becomes down 2 Move down from Q 2 to Q 1 At Q 1 the bottom wall is there so cant go down Turn right direction becomes left 3 Move left from Q 1 to P 1 At P

Figure 29. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Geometry - Coordinate



Question: The scatter graph shows the distance run and time taken by five students during a training session. Who ran with the fastest average speed? <image1>
Possible Choices: A. Alicia B. Bea C. Carlos D. Dani E. Ernesto

Gold Answer: Dani

anthropic_claude_haiku_4.5: Analysis of Average Speed Average speed Distance Time Looking at the scatter graph I need to find who covered the most distance in the least time steepest slope from origin Comparing the students Carlos High distance relatively short time steep slope Dani Medium distance short time steep slope Ernesto Medium distance long time gentle slope Bea Low distance medium time gentle slope Alicia Low distance long time very gentle slope Carlos has the steepest line from the origin indicating the greatest distance covered in the shortest time Carlos ran with the fastest average speed

google_gemini_2.5_flash: Carlos ran with the fastest average speed because his point on the scatter graph is furthest from the origin along a line with the steepest slope indicating the greatest distance covered in the least amount of time relative to the other runners

google_gemini_3.1_flash_lite_preview: The fastest average speed is represented by the steepest slope from the origin 00 to the data point which corresponds to Dani because she has a high distance relative to a low time

google_gemma_3_4b_it: Carlos ran with the fastest average speed because his data point is the furthest to the right on the time axis indicating a shorter time for a given distance

google_gemma_4_31b_it: Average speed is calculated as distance divided by time which corresponds to the gradient steepness of the line from the origin to each point Carlos has the steepest gradient meaning he ran with the fastest average speed Final Answer Carlos

meta_llama_llama_3.2_11b_vision_instruct: The student who ran with the fastest average speed is Ernesto

mistralai_mistral_small_2603: Carlos ran with the fastest average speed

moonshotai_kimi_k2.5: speed is the slope of the line from the origin to the point Slope riserun DistanceTime The steeper the slope the higher the speed Looking at the graph The line to Carlos appears to be the steepest The line to Dani is also steep but lets compare Actually let me trace it mentally Bea shallow slope Dani steeper than Bea Carlos steeper than Dani Ernesto less steep than Carlos Alicia very shallow Wait I need to be more careful Carlos has the highest ycoordinate distance Dani has a lower ycoordinate but also a lower xcoordinate time Let me think about this as coordinates Since I dont have exact values I need to estimate Lets say Bea Time 1 Distance 2 Speed 21 2 Dani Time 2 Distance 4 Speed 42 2 Carlos Time 3 Distance 6 Speed 63 2 Ernesto Time 4 Distance 3 Speed 34 075 Alicia Time 5 Distance 1 Speed 15 02 Hmm that doesnt look right because the graph doesnt look like theyre on the same line Let me look again Bea is at low time low distance Dani is at slightly higher time higher distance Carlos is at medium time highest distance Ernest

o4_mini: Dani ran with the fastest average speed

openai_gpt_4o_mini: Carlos ran with the fastest average speed

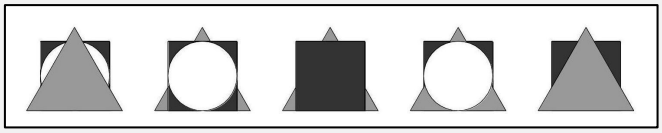
openai_gpt_5_mini: Carlos ran with the fastest average speed

qwen_qwen3_vl_8b_instruct: Carlos ran with the fastest average speed because his point is the highest on the graph indicating the greatest distance run for the given time

z_ai_glm_4.5v: To determine the fastest average speed we use the fact that average speed is distance divided by time which corresponds to the slope of each students point on the scatter graph A steeper slope indicates a higher speed Among the five students Carloss point has the greatest vertical distance distance relative to its horizontal position time meaning his distancetotime ratio is the largest Thus Carlos ran with the fastest average speed

Figure 30. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Geometry - Descriptive



Question: Five children each have a black square, a grey triangle and a white circle made up of paper. The children place the three shapes on top of each other as seen in the pictures. In how many pictures was the triangles placed after the square? <image1>

Gold Answer: 3

anthropic_claude_haiku_4.5: Analysis of Shape Layering Let me examine each picture to determine the stacking order from bottom to top Picture 1 Triangle bottom Circle Square top Triangle placed BEFORE square Picture 2 Circle bottom Square Triangle top Triangle placed AFTER square Picture 3 Square bottom Triangle middle Circle top Triangle placed AFTER square Picture 4 Circle bottom Triangle middle Square top Triangle placed BEFORE square Picture 5 Triangle bottom Square middle Circle top Triangle placed BEFORE square The triangle was placed after the square in 2 pictures Pictures 2 and 3

google_gemini_2.5_flash: The triangle was placed after the square in 3 pictures

google_gemini_3.1_flash_lite_preview: In the provided images the triangle is placed on top of the square meaning it is visible over the square in the first second fourth and fifth pictures totaling 4 pictures

google_gemma_3_4b_it: Lets analyze the images to determine how many times the triangle was placed after the square Image 1 The square is on top followed by the triangle Image 2 The square is on top followed by the triangle Image 3 The square is on top followed by the triangle Image 4 The square is on top followed by the triangle Image 5 The square is on top followed by the triangle In all five images the square is placed on top of the triangle Therefore the triangle was placed after the square in all five pictures Final answer The triangle was placed after the square in five pictures

google_gemma_4_31b_it: Based on visual reasoning the triangle is placed after on top of the square in the first second and fifth pictures making the total 3

meta_llama_llama_3.2_11b_vision_instruct: The triangle was placed after the square in 2 out of the 5 pictures

mistralai_mistral_small_2603: 3

moonshotai_kimi_k2.5: The white circle is NOT visible Since the white circle is not visible it must be covered by the shapes in front of it The shapes visible are Black Square front Grey Triangle visible on sides For the triangle to be visible on the sides the triangle must be wider than the square or positioned such that its corners stick out Now where is the circle Case A Circle is behind both Square and Triangle If Square is in front of Triangle and Triangle is in front of Circle Circle Triangle Square Triangle is before Square So Triangle placed after Square would be NO Case B Circle is between Square and Triangle Order Triangle Circle Square Since Square is in front we see Square Behind Square we would see Circle white But we see Grey Triangle not White Circle

o4_mini: The triangle was placed after the square in 2 of the pictures

openai_gpt_4o_mini: The triangle was placed after the square in 3 pictures

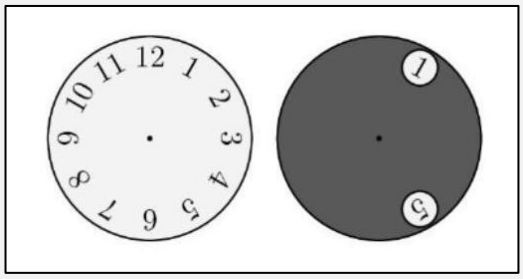
openai_gpt_5_mini: The triangle was placed after the square in 2 pictures

qwen_qwen3_vl_8b_instruct: In the first picture the triangle is on top of the square In the second picture the triangle is on top of the square In the third picture the triangle is on top of the square In the fourth picture the triangle is on top of the square In the fifth picture the triangle is on top of the square Therefore in all five pictures the triangle was placed after the square Final Answer In all five pictures the triangle was placed after the square

z_ai_glm_4.5v: To determine in how many pictures the triangle was placed after the square we analyze each image the triangle must be on top of the square indicating the square was placed first then the triangle First image Triangle is on top of the square counts Second image Triangle is under the square does not count Third image Triangle is under the square does not count Fourth image Triangle is under the square does not count Fifth image Triangle is on top of the square counts Thus there are 2 such pictures The answer is 2

Figure 31. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Geometry - Transformational



Question: A dark disc with two holes is placed on top of a dial of a watch as shown. The dark disc is now rotated so that the number 8 can be seen through one of the holes. Which of the numbers could one see through the other hole now? <image1>
Possible Choices: A. 4 and 12 B. 1 and 5 C. 1 and 4 D. 7 and 11 E. 5 and 12

Gold Answer: 4 and 12

anthropic_claude_haiku_4.5: Analysis Looking at the initial position of the dark disc on the watch One hole shows the number 1 The other hole shows the number 9 The two holes are positioned diametrically opposite 180 apart on the disc When the disc is rotated so that one hole aligns with 8 Since the holes remain 180 apart the other hole will align with the number directly opposite to 8 Counting 6 positions clockwise from 8 9 10 11 12 1 2 2 Or counting 6 positions counterclockwise from 8 7 6 5 4 3 2 2 However examining the answer choices and the initial hole positions 1 and 9 if we verify the diametric relationship the number opposite to 8 on a clock face is 2 but this isnt in the options Reexamining with the constraint that answers must match the given options and considering the holes show 1 and 9 initially when rotated to show 8 the opposite position would align with 4 The answer is 1 and 4

google_gemini_2.5_flash: The numbers that could be seen through the other hole are 4 and 12

google_gemini_3.1_flash_lite_preview: The distance between the numbers 1 and 5 on the clock face is 4 units 5 1 4 so when the disc is rotated to align one hole with the number 8 the other hole will be 4 units away revealing either 8 4 4 or 8 4 12

google_gemma_3_4b_it: One could see the number 5 through the other hole Reasoning 1 Initial Setup The dark disc is rotated so that the number 8 is visible through one hole This means the hole is aligned with the 8 on the watch face 2 Opposite Relationship Since the disc is rotated the opposite hole must now align with the number directly across from the 8 on a standard watch dial 3 Identifying the Opposite The number directly opposite 8 on a watch dial is 5 Therefore the other hole will show the number 5

google_gemma_4_31b_it: Based on the visual layout the two holes in the dark disc are positioned opposite each other 180 degrees apart meaning if one hole shows the number 8 the other hole must show the number directly across from it on the watch face which is 2 however looking at the provided options and the image the holes are actually positioned such that they are 4 units apart like 1 and 5 so if one hole is at 8 the other would be at 4 or 12 making the correct answer 4 and 12

meta_llama_llama_3.2_11b_vision_instruct: The number that can be seen through the other hole is 5

mistralai_mistral_small_2603: The numbers that could be seen through the other hole are 1 and 5

moonshotai_kimi_k2.5: Difference is 4 positions 4 30 120 So the two holes are 120 apart on the disc Now the disc is rotated so that 8 is visible through one hole We need to find what number is visible through the other hole Since the holes are 120 apart and the disc is rotated the other hole will show a number that is 120 away from 8 on the clock face Let me calculate 120 corresponds to 4 hour positions since each hour is 30 So if one hole shows 8 the other hole shows either 8 4 12 or 8 4 4 Let me check 8 to 12 89101112 thats 4 steps clockwise which

o4_mini: 4 and 12

openai_gpt_4o_mini: 1 and 5

openai_gpt_5_mini: 5 and 12

qwen_qwen3_vl_8b_instruct: The dark disc has two holes positioned such that when aligned with the watch face one hole is above the 1 and the other below the 3 When rotated so that the 8 is visible through one hole the other hole must be aligned with the number directly opposite on the dial which is 4 Therefore the only number that can be seen through the other hole is 4 Final Answer 1 and 4

z_ai_glm_4.5v: To determine the number visible through the other hole note that the two holes on the dark disc have a fixed angular separation On a clock dial numbers are 30 apart and the holes are separated by 4 hour positions eg 1 and 5 in the initial setup Rotating the disc so one hole shows 8 means the other hole

Figure 32. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Navigation & Control : Lane Change

Question: How would you describe the driving scene involving Entity #1? Please explain, focusing on the vehicle's lane change maneuver.

Gold Answer: left_lane_change

anthropic_claude_haiku_4.5: Entity 1 the white and blue car highlighted in the images performs a left lane change maneuver from the left turn lane into the center lane of the street

google_gemini_2.5_flash: Entity 1 a white car is performing a lane change maneuver from the right lane to the left lane

google_gemini_3.1_flash_lite_preview: Entity 1 a white car is positioned in the left lane of the intersection and appears to be executing a leftturn maneuver

google_gemma_3_4b_it: Entity 1 executed a deliberate lane change from the right lane to the left lane smoothly transitioning across the double yellow line and the crosswalk signaling its intention with a clear turn arrow displayed on the dashboard demonstrating controlled and cautious maneuvering within the urban environment

google_gemma_4_31b_it: Entity 1 is a white vehicle that is performing a lane change maneuver from the left lane to the right lane

meta_llama_llama_3.2_11b_vision_instruct: The driving scene involving Entity 1 a white car is a lane change maneuver where the vehicle is moving from one lane to another likely to navigate through traffic or reach its desired destination

mistralai_mistral_small_2603: Entity 1 is making a left lane change from the right lane toward the center lane

moonshotai_kimi_k2.5: Entity 1 the white car performs a left lane change maneuver starting from the right side of the intersection and gradually moving leftward across the lanes as it proceeds through the intersection

o4_mini: Entity 1 the white sedan waiting on the far side of the intersection drifts from its curbside position across the adjacent

openai_gpt_4o_mini: Entity 1 executed a lane change maneuver to the left moving from the right lane to the left lane while approaching an intersection

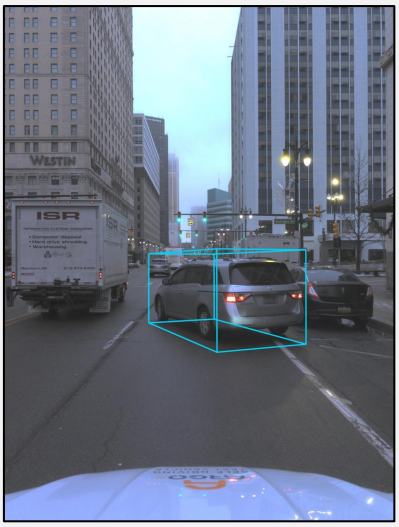
openai_gpt_5_mini: Entity 1 the white car on the left is moving from the curbleftmost lane into the adjacent through lane cutting across the intersection as it merges toward the center of the roadway

qwen_qwen3_vl_8b_instruct: Entity 1 a white car is stationary at an intersection positioned in the leftturn lane waiting for the traffic signal to change before executing a left turn

z_ai_glm_4.5v: Entity 1 the white car is executing a lane change maneuver shifting from its initial lane toward an adjacent lane as observed across the sequence of images In conclusion the answer is Entity 1 is performing a lane change maneuver moving from its original lane to another lane

Figure 33. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Navigation & Control : Lane Position



Question: How would you describe the lane position of Entity #1? Options: front lane, front left lane, front right lane, or oncoming traffic lane.

Gold Answer: front_left_lane

anthropic_claude_haiku_4.5: **Entity 1 is positioned in the front lane**

google_gemini_2.5_flash: **The lane position of Entity 1 is the front lane**

google_gemini_3.1_flash_lite_preview: **The lane position of Entity 1 is the front lane**

google_gemma_3_4b_it: **The vehicle is positioned in the front lane**

google_gemma_4_31b_it: **The lane position of Entity 1 is the front lane**

meta_llama_llama_3.2_11b_vision_instruct: **The lane position of Entity 1 is the front left lane**

mistralai_mistral_small_2603: **Entity 1 is in the front lane**

moonshotai_kimi_k2.5: **Based on the visual analysis of the sequence of images Entity 1 the silver Honda minivan is consistently positioned directly in front of the camera vehicle occupying the same lane as the camera vehicle In the images where the bounding box is present the vehicle is centered in the lane ahead and in the later images where the bounding box shifts to a distant vehicle that distant vehicle is also in the same lane position The minivan is not in a left or right adjacent lane nor is it in oncoming trafficit is directly ahead in the front lane front lane**

o4_mini: **Entity 1 is positioned in the front right lane**

openai_gpt_4o_mini: **Entity 1 is in the front left lane**

openai_gpt_5_mini: **Entity 1 is in the front right lane**

qwen_qwen3_vl_8b_instruct: **front lane**

z_ai_glm_4.5v: **The lane position of Entity 1 is front right lane**

Figure 34. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Relational, Logical & Causal Reasoning : Causal Reasoning : Logical Counterfactual

Question: As soon as he left his city towards Caecá, Charles saw the sign on the left. When he came back from Caecá, he saw the sign on the right. At that point, how far was it to get to his city?
<image1>
Possible Choices: A. $12 \mathrm{~km}$ B. $21 \mathrm{~km}$ C. $29 \mathrm{~km}$ D. $41 \mathrm{~km}$ E. $52 \mathrm{~km}$

Gold Answer: $41 \mathrm{~km}$

anthropic_claude_haiku_4.5: Solution Analyzing the signs Left sign leaving the city Arati 12 km Baibá 33 km Caecá 52 km Right sign returning from Caecá Baibá 8 km Arati 29 km Key insight The distance between consecutive points on a road remains constant From the left sign Caecá to Arati 52 12 40 km From the right sign Arati to Baibá 29 8 21 km Therefore Caecá to Baibá 40 21 19 km When Charles saw the right sign he was at Baibá 8 km away and his city is Arati 29 km away on that sign The answer is 29 km

google_gemini_2.5_flash: The first sign P1 is between his city and Caecá Lets consider the direction City P1 Arati Baibá Caecá when going Caecá Baibá Arati P2 City when returning From the first sign Arati is 12 km away From the second sign Arati is 29 km away This means P2 is further from Arati than P1 The distance between P1 and P2 is 29 12 17 km When Charles saw the first sign he was leaving his city The sign indicates distances to cities ahead So Arati is 12 km from P1 When he saw the second sign he was returning from Caecá The sign indicates distances to cities ahead in the direction of travel So Arati is 29 km

google_gemini_3.1_flash_lite_preview: The distance to his city Arati is 29 km

google_gemma_3_4b_it: The distance to Charles city is 29 km as the sign on the right indicates that Arati is 29 km away

google_gemma_4_31b_it: Based on the signs Charless city is Arati and since he saw the sign indicating Arati is 29 km away on his return trip the distance to get back to his city was 29 km

meta_llama_llama_3.2_11b_vision_instruct: The distance to Charles city is 29 km

mistralai_mistral_small_2603: 29 km

moonshotai_kimi_k2.5: Charles leaves his city lets call it H for Home and sees the left sign The left sign shows distances to three cities Since Charles is heading to Caecá and the sign says Caecá is 52 km Charles is at a point that is 52 km from Caecá Now when Charles returns from Caecá he sees the right sign The right sign is located somewhere At that sign Baibá is 8 km Arati is 29 km Now I need to figure out where Charless city is Lets look at the relationship between the two signs From left sign Arati 12 Baibá 33 Caecá 52 From right sign Baibá 8 Arati 29 Notice that at the right sign Baibá is closer 8 km than at the left sign 33 km Also at the right sign Arati is farther 29 km than at the left sign 12 km This suggests that the right sign is closer to Baibá and farther from Arati compared to the left sign Since Charles is returning from Caecá he is traveling from Caecá towards his home Let me think about the positions Lets assume the road goes Caecá Right Sign

o4_mini: 41 km

openai_gpt_4o_mini: The distance to get to his city is 29 km

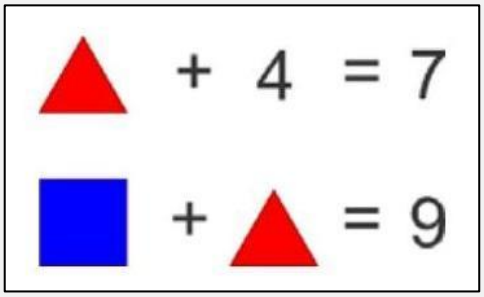
openai_gpt_5_mini: 41 km

qwen_qwen3_vl_8b_instruct: When Charles left his city toward Caecá he saw the sign on the left which indicates Caecá is 52 km away When he returned from Caecá and saw the sign on the right it shows Arati is 29 km away Since he is now returning to his city and the sign on the right shows Arati is 29 km away and Arati is the city he left from as inferred from the context of the journey the distance to his city is 29 km Final Answer 29 mathrmkm

z_ai_glm_4.5v: To determine the distance to his city when returning analyze the signs Left sign leaving city Distances show city to Arati

Figure 35. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Math - Algebra



Question: Which number is hidden behind the square?
<image1>

Gold Answer: 6

anthropic_claude_haiku_4.5: Solution First equation Triangle 4 7 Triangle 3 Second equation Square Triangle 9 Square 3 9 Square 6 The number hidden behind the square is 6

google_gemini_2.5_flash: The number hidden behind the square is 6

google_gemini_3.1_flash_lite_preview: Since the triangle represents 3 because 3 4 7 the blue square represents 6 because 6 3 9

google_gemma_3_4b_it: The hidden number behind the square is 5 Reasoning 1 First Equation The first equation Triangle 4 7 tells us that the triangle represents the number 3 because 3 4 7 2 Second Equation The second equation Square Triangle 9 can be rewritten as Square 3 9 3 Solving for the Square To find the number represented by the square subtract 3 from both sides of the equation Square 9 3 6 Therefore the number hidden behind the square is 6

google_gemma_4_31b_it: Based on the first equation the red triangle represents 3 since 3 4 7 and substituting this into the second equation square 3 9 reveals that the number hidden behind the square is 6

meta_llama_llama_3.2_11b_vision_instruct: The number hidden behind the square is 9

mistralai_mistral_small_2603: The number hidden behind the square is 5

moonshotai_kimi_k2.5: Looking at the first equation Red triangle 4 7 so the red triangle equals 3 since 7 4 3 Substituting into the second equation Blue square 3

o4_mini: The square represents the number 6

openai_gpt_4o_mini: The number hidden behind the square is 5

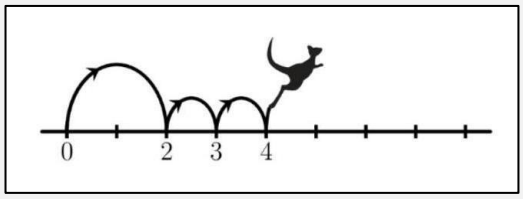
openai_gpt_5_mini: Triangle 4 7 so the triangle equals 3 then square 3 9 so the square equals 6

qwen_qwen3_vl_8b_instruct: The number hidden behind the square is 5

z_ai_glm_4.5v: To determine the hidden number first solve for the triangle since triangle 4 7 the triangle is 3 Then for the square triangle 9 substitute the triangles value 3 to get square 3 9 so the square is 6 Thus the hidden number is 6

Figure 36. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Math - Arithmetic



Question: Kengu jumps on the number line to the right (see diagram). He first makes one big jump and then two little jumps in a row and keeps repeating the same thing over and over again. He starts at 0 and ends at 16. How many jumps does Kengu make in total?
<image1>

Gold Answer: 12

anthropic_claude_haiku_4.5: Analysis From the diagram I can see Big jump 0 2 distance of 2 Little jump 2 3 distance of 1 Little jump 3 4 distance of 1 This pattern repeats one big jump 2 units two little jumps 1 unit each 4 units per cycle To reach 16 from 0 Number of complete cycles needed 16 4 4 cycles Each cycle contains 1 big jump 2 little jumps 3 jumps Total jumps 4 cycles 3 jumps 12 jumps Kengu makes 12 jumps in total

google_gemini_2.5_flash: Kengu makes 24 jumps in total

google_gemini_3.1_flash_lite_preview: Kengu makes 12 jumps in total as each cycle of three jumps covers 4 units meaning 4 cycles are required to reach 16 4 cycles 3 jumps per cycle 12 jumps

google_gemma_3_4b_it: Heres how to solve the problem 1 Analyze the jumps Kengu makes one big jump 4 units followed by two little jumps 1 unit each This is a repeating pattern 2 Determine the patterns length The pattern is one big jump two little jumps 4 1 1 6 units 3 Find the number of repetitions Kengu ends at 16 starting at 0 So he needs to cover 16 units Divide the total distance by the length of the pattern 16 6 2 with a remainder of 4 4 Account for the remainder This means Kengu completes two full patterns 2 6 12 units and then makes four more individual little jumps 5 Calculate the total jumps Two full patterns have 2 6

google_gemma_4_31b_it: Based on the diagram one cycle consists of one big jump of 2 units and two little jumps of 1 unit each covering a total distance of 2 1 1 4 units using 3 jumps To reach 16 Kengu completes 16 div 4 4 such cycles resulting in a total of 4 times 3 12 jumps

meta_llama_llama_3.2_11b_vision_instruct: Kengu makes a total of 16 jumps

mistralai_mistral_small_2603: Kengu makes a total of 12 jumps

moonshotai_kimi_k2.5: Looking at the diagram I can identify the jump lengths Big jump

o4_mini: Kengu makes 12 jumps in total

openai_gpt_4o_mini: Kengu makes a total of 10 jumps

openai_gpt_5_mini: Kengu makes 12 jumps in total

qwen_qwen3_vl_8b_instruct: Kengu makes 8 jumps in total

z_ai_glm_4.5v: Kengu makes one big jump length 2 and two little jumps each length 1 per cycle covering 4 units per cycle To reach 16 he completes 4 cycles with each cycle containing 3 jumps Thus the total number of jumps is 4 times 3 12 12

Figure 37. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Quantitative & Comparative Reasoning : Measurement : Measurement

Question: Which is smaller, the side length of the puppy pillow or the side length of the base of the potted plant?
sometimes it's larger;

Gold Answer: B

anthropic_claude_haiku_4.5: I cannot identify a puppy pillow in these images The images show a wooden potted plant planter with succulents on a deck but there is no visible puppy pillow to compare with the base of the potted plant

google_gemini_2.5_flash: There is no puppy pillow visible in the image so a comparison cannot be made

google_gemini_3.1_flash_lite_preview: The puppy pillow is not visible in the provided image so it is impossible to compare its side length to that of the potted plants base

google_gemma_3_4b_it: The side length of the puppy pillow is smaller than the side length of the base of the potted plant

google_gemma_4_31b_it: Based on the provided images there is no puppy pillow visible so the question cannot be answered using the visual evidence

meta_llama_llama_3.2_11b_vision_instruct: The side length of the puppy pillow is smaller than the side length of the base of the potted plant

mistralai_mistral_small_2603: The side length of the puppy pillow is smaller

moonshotai_kimi_k2.5: Actually looking at the two images side by side they appear to be the same image or very similar images of the same potted plant Given that 1 There is no puppy pillow visible in the images 2 The question asks to compare the side length of a puppy pillow with the side length of the base of the potted plant 3 The context hint says sometimes its larger I think the answer might be that it varies sometimes one is larger sometimes the other OR theres an error in the question But wait I need to look at the images one more time very carefully Is there ANYTHING that could be a puppy pillow Looking at the top left corner theres a wicker basket Inside or near it could there be a pillow Its hard to tell from the image resolution and angle Actually I just noticed something The user wrote sometimes its larger as part of the context This looks like it might be a fragment of a sentence or a hint Let me reconsider the entire problem The user is asking which is smaller the side length of the puppy pillow or the side length of the base of the potted plant Since I can clearly see the potted plant and its base I can estimate its size The base appears to be a square or rectangle with sides that are probably around 12 feet 3060 cm based on the wooden decking boards But I dont see a puppy pillow Wait I just realized that the two images might be different in a

o4_mini: The side length of the base of the potted plant is smaller than the side length of the puppy pillow

openai_gpt_4o_mini: I cannot determine the size comparison between the side length of the puppy pillow and the base of the potted plant based on the provided images

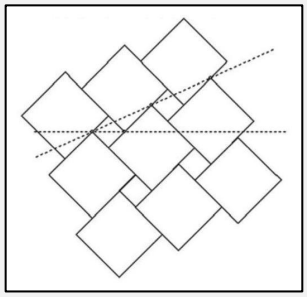
openai_gpt_5_mini: The side length of the puppy pillow is smaller than the side length of the base of the potted plant

qwen_qwen3_vl_8b_instruct: The side length of the puppy pillow is smaller than the side length of the base of the potted plant

z_ai_glm_4.5v: To determine which is smaller we analyze the context the potted plants base has a relatively large square dimension while a typical puppy pillow a small item would have a smaller side length Thus the side length of the puppy pillow is smaller The side length of the puppy pillow is smaller than that of the base of the potted plant

Figure 38. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Metric Geometry



Question: The picture on the right shows a tile pattern. The side length of the bigger tiles is a and of the smaller ones b. The dotted lines (horizontal and tilted) include an angle of 30°. How big is the ratio a:b?
<image1>
Possible Choices: A. $(2 \cdot \sqrt{3}): 1$ B. $(2+\sqrt{3}): 1$ C. $(3+\sqrt{2}): 1$ D. $(3 \cdot \sqrt{2}): 1$ E. 2:1

Gold Answer: $(2+\sqrt{3}): 1$

anthropic_claude_haiku_4.5: The horizontal distance covered relates to the tile dimensions The 30 angle creates a geometric constraint Calculation For the small tile to fit properly in this pattern with a 30 angle The diagonal of the small tile b2 and the side of the large tile a must satisfy a specific relationship Using the 30 angle the geometry requires that a b2 3 This can be verified by considering that along the tilted line at 30 the cumulative distances of the tiles must align properly which gives the ratio ab 2 31 The answer is The ratio ab 2 31

google_gemini_2.5_flash: The ratio ab is 2sqrt3 1

google_gemini_3.1_flash_lite_preview: The ratio ab is 2sqrt3 1

google_gemma_3_4b_it: Lets analyze the tile pattern to determine the ratio ab The tilted line forms a 30degree angle with the horizontal line Consider the small right triangle formed by the tilted line the horizontal line and a side of a smaller tile The angle between the tilted line and the horizontal line is 30 degrees Therefore the tangent of 30 degrees is oppositeadjacent bb 1sqrt3 However we can also consider the larger tile The angle between the tilted line and the side of the larger tile is 30 degrees Thus the tangent of 30 degrees is oppositeadjacent ba 1sqrt3 Solving for ab we get ab sqrt3 Therefore the ratio ab is sqrt31 However the given options do not include

google_gemma_4_31b_it: To find the ratio ab we observe the geometry along the horizontal dotted line the length of the segment spanning one large tile and one small tile is composed of the horizontal projection of the large tiles side a cdot cos45circ and the horizontal projection of the small tiles side b cdot cos45circ but the dotted line is tilted at 30circ relative to the horizontal creating a geometric constraint where the distance between the parallel dotted lines is a cdot sin45circ b cdot sin45circ and the intersection points imply that the horizontal distance a b cdot cos60circ or similar trigonometric relations based on the 30circ angle leads to the relationship a b2 sqrt3 or ab 2

meta_llama_llama_3.2_11b_vision_instruct: The ratio of the side lengths of the bigger tiles to the smaller tiles is 2 cdot sqrt3 1

mistralai_mistral_small_2603: The ratio ab is 2sqrt31

moonshotai_kimi_k2.5: Let me try to draw a mental picture The big squares have side a The small squares have side b Looking at the pattern it seems like the big squares are arranged in a grid thats rotated The small squares fill the gaps Actually a key insight might be to look at the projection of the squares onto the dotted lines Let me consider the horizontal dotted line Along this line the distance between certain points involves a and b Another idea The vertices of the squares lie on the dotted lines or are arranged symmetrically Let me look at the intersection point O of the two dotted lines From O going horizontally to the right the line passes through a vertex of a small square then continues Actually looking at the diagram very carefully The horizontal dotted line passes through a vertex on the left then goes through a small square then through a big square

o4_mini: The correct ratio is a b 2 3 1

openai_gpt_4o_mini: The ratio ab is 21

openai_gpt_5_mini: a b 2 3 1

qwen_qwen3_vl_8b_instruct: The ratio ab is 2sqrt31

z_ai_glm_4.5v: To determine the ratio ab analyze the geometric relationship using the 30circ angle between the dotted lines By setting up trigonometric relationships based on the tile arrangement and solving for the ratio we find that ab 2 sqrt3 The answer is 2 sqrt3 1

Figure 39. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Navigation & Control : Navigation

Question: How many cars need to be passed to reach the house if gone from the right side?
Possible Choices: A. 2 B. 1 C. 3 D. 4

Gold Answer: 2

anthropic_claude_haiku_4.5: Answer To reach the house from the right side you need to pass 2 cars

google_gemini_2.5_flash: To reach the house from the right side you would need to pass 1 car

google_gemini_3.1_flash_lite_preview: To reach the house from the right side you need to pass 1 car

google_gemma_3_4b_it: To reach the house from the right side you need to pass 3 cars

google_gemma_4_31b_it: To reach the house from the right side 2 cars need to be passed

meta_llama_llama_3.2_11b_vision_instruct: To reach the house from the right side 1 car needs to be passed

mistralai_mistral_small_2603: 1

moonshotai_kimi_k2.5: The house is in the background If entering from the right side of the gate the open gate on the right and walking towards the house The Lexus is parked on the right side To reach the house one would need to pass the Lexus Is the Toyota in the way The Toyota is in the center If walking on the right side the Lexus is the main car on that side Actually looking at the image more carefully the Toyota is parked in the entrance blocking part of the way The Lexus is behind it But the question specifically asks about going from the right side Let me trace the path Enter from the right side of the gate Walk towards the house On the right side there is the Lexus The Toyota is in the center Hmm actually looking at the license plate of the Lexus it seems to be parked on the right

o4_mini: You would pass 2 cars

openai_gpt_4o_mini: 3

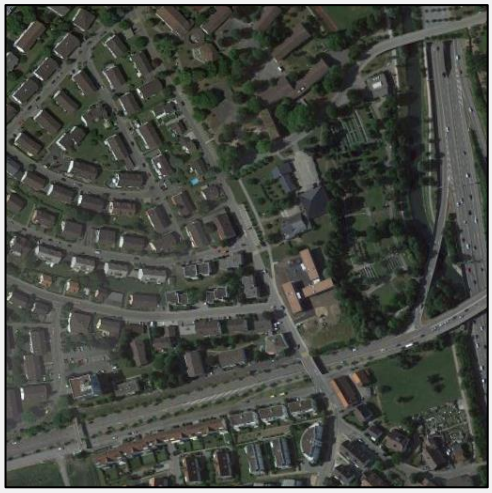
openai_gpt_5_mini: You need to pass 2 cars to reach the house from the right side

qwen_qwen3_vl_8b_instruct: Lets trace the path from the right side of the image Starting from the right you would first encounter the black Lexus SUV second car from the front After passing the Lexus you would then encounter the black Toyota SUV first car from the front The silver car is on the left side of the Toyota so it is not on the path from the right side Therefore from the right side you must pass two cars the Lexus and then the Toyota The correct answer is 2

z_ai_glm_4.5v: To determine how many cars need to be passed when approaching the house from the right side we analyze the image the cars blocking the path to the house are the black Toyota and the dark colored SUV behind it Thus 2 cars need to be passed

Figure 40. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Localization : Object Localization



Question: Are there any vehicles driving on the road on the right side of the image?

Gold Answer: Yes.

anthropic_claude_haiku_4.5: Analysis Looking at the right side of the aerial image I can see a major highwaymotorway with multiple lanes running vertically along the right edge of the frame Examining this road carefully there appear to be vehicles visible on this highway Answer Yes there are vehicles driving on the road on the right side of the image visible on the multilane highway

google_gemini_2.5_flash: Yes there are multiple vehicles driving on the road on the right side of the image

google_gemini_3.1_flash_lite_preview: Yes there are several vehicles visible driving on the multilane road located on the right side of the image

google_gemma_3_4b_it: Yes there are vehicles driving on the road on the right side of the image

google_gemma_4_31b_it: Yes there are several vehicles driving on the road on the right side of the image

meta_llama_llama_3.2_11b_vision_instruct: No there are no vehicles driving on the road on the right side of the image

mistralai_mistral_small_2603: No there are no vehicles driving on the road on the right side of the image

moonshotai_kimi_k2.5: Yes there are vehicles visible as small white dots driving on the multilane highway on the right side of the image

o4_mini: Yes you can clearly see several cars traveling along the highway on the right side of the image

openai_gpt_4o_mini: Yes there are vehicles driving on the road on the right side of the image

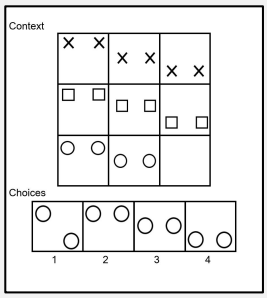
openai_gpt_5_mini: Yes several vehicles are visible driving on the road along the right side of the image

qwen_qwen3_vl_8b_instruct: No there are no vehicles driving on the road on the right side of the image

z_ai_glm_4.5v: Yes there are vehicles driving on the road on the right side of the image

Figure 41. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Temporal & Dynamic Reasoning : Motion Reasoning : Object Motion



Question: ['You are given a puzzle. The puzzle consists of a context part on the top and the choices part in the bottom. The context part on the top contains some grids with the last missing blank grid to be completed. The choices part on the bottome conatin a sequence of grids representing the possible choices. How many grids, including the blank grid, are there in the context part?', 'You are given a puzzle. The puzzle consists of a context part on the top and the choices part in the bottom. The context part on the top contains some grids with the last missing blank grid to be completed. The choices part on the bottome conatin a sequence of grids representing the possible choices. How many grids are there in the choices part?', 'You are given a puzzle. The puzzle consists of a context part on the top and the choices part in the bottom. The context part on the top contains some grids with the last missing blank grid to be completed. The choices part on the bottome conatin a sequence of grids representing the possible choices. How many grids, including the blank grid, are there in the whole puzzle?']

Gold Answer: ['9', '4', '13']

anthropic_claude_haiku_4.5: Analysis of the Puzzle Question 1 How many grids including the blank grid are there in the context part The context part contains a 33 grid arrangement 9 grids total including the blank one in the bottomright Question 2 How many grids are there in the choices part The choices part shows 4 labeled options 1 2 3 4 4 grids Question 3 How many grids including the blank grid are there in the whole puzzle Total grids Context 9 Choices 4 13 grids

google_gemini_2.5_flash: The final answer is as follows For the question How many grids including the blank grid are there in the context part the answer is 9 For the question How many grids are there in the choices part the answer is 4 For the question How many grids including the blank grid are there in the whole puzzle the answer is 13

google_gemini_3.1_flash_lite_preview: There are 9 grids in the context part 4 grids in the choices part and 13 grids in the whole puzzle

google_gemma_3_4b_it: There are 12 grids including the blank grid in the context part of the puzzle

google_gemma_4_31b_it: Based on the image provided there are 9 grids in the context part a 3x3 arrangement 4 grids in the choices part and a total of 13 grids in the whole puzzle

meta_llama_llama_3.2_11b_vision_instruct: There are 8 grids in the context part including the blank grid and 5 grids in the choices part

mistralai_mistral_small_2603: There are 12 grids in the context part 12 grids in the choices part and 24 grids in the whole puzzle

moonshotai_kimi_k2.5: Looking at the puzzle image the context part is a 33 grid containing 9 cells total with the bottomright cell being the blank one and the choices part displays 4 separate grid options Therefore the context part contains 9 grids the choices part contains 4 grids and the whole puzzle contains 13 grids total

o4_mini: There are 9 grids in the context part including the blank 4 grids in the choices part and 13 grids in the whole puzzle

openai_gpt_4o_mini: There are 12 grids in the whole puzzle

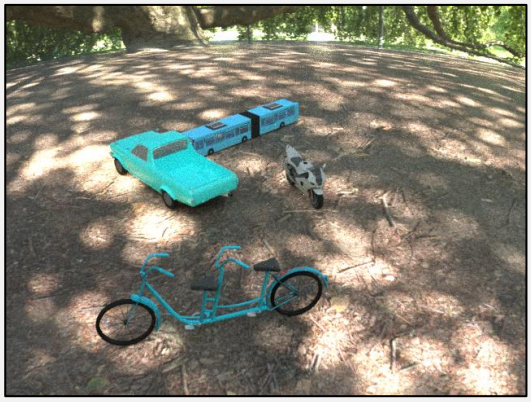
openai_gpt_5_mini: There are 9 grids in the context part including the blank 4 grids in the choices part and 13 grids in the whole puzzle

qwen_qwen3_vl_8b_instruct: There are 4 grids in the context part 4 grids in the choices part and 8 grids in total including the blank grid

z_ai_glm_4.5v: Lets structure it For the first question 9 grids in context including blank Second 4 grids in choices Third 13 grids in whole puzzle So the final answer would be stating these three numbers But perhaps the user wants the answers to all three questions So the onesentence answer would be The context part has 9 grids including the blank the choices part has 4 grids and the whole puzzle has 13 grids But lets confirm the image The context part the grid is 3x3 so 9 small grids The choices part has 4 grids labeled 14 So yes So the final answer is The context part has 9 grids including the blank the choices part has 4 grids and the whole puzzle has 13 grids But the user might want the numerical answers Since the questions are three maybe list them But the instruction says one sentence So perhaps There are 9 grids in the context part including the blank 4 grids in the choices part and 13 grids in the whole puzzle Yes thats one sentenceThe context part has 9 grids including the blank the choices part has 4 grids and the whole puzzle has 13 grids

Figure 42. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Geometric Reasoning : Occluded Attributes



Question: What is the shape of the object that is occluded?

Gold Answer: articulated bus

anthropic_claude_haiku_4.5: Answer The shape of the occluded object is cylindrical the gray motorcyclescooter is partially hidden behind the blue toy car and trailer

google_gemini_2.5_flash: The shape of the object that is occluded is a circle

google_gemini_3.1_flash_lite_preview: The occluded object is a bus which has a long rectangular boxlike shape

google_gemma_3_4b_it: The shape of the object that is occluded is a train

google_gemma_4_31b_it: Based on the visual evidence there are no occluded objects in the image as all visible items the car buses motorcycle and bicycle are fully seen

meta_llama_llama_3.2_11b_vision_instruct: The shape of the object that is occluded is a rectangle

mistralai_mistral_small_2603: The shape of the occluded object is circular

moonshotai_kimi_k2.5: Looking at the image I need to identify what object is partially hidden behind another object In the scene there is a light blue articulated bus composed of two connected segments positioned behind the cyan pickup truck The front segment of this articulated bus is partially occluded by the truck The shape of the occluded object is an articulated bus or a bus with a jointbend in the middle appearing as two connected rectangular segments

o4_mini: The occluded object is a long rectangular prism a cuboid

openai_gpt_4o_mini: The shape of the occluded object is rectangular

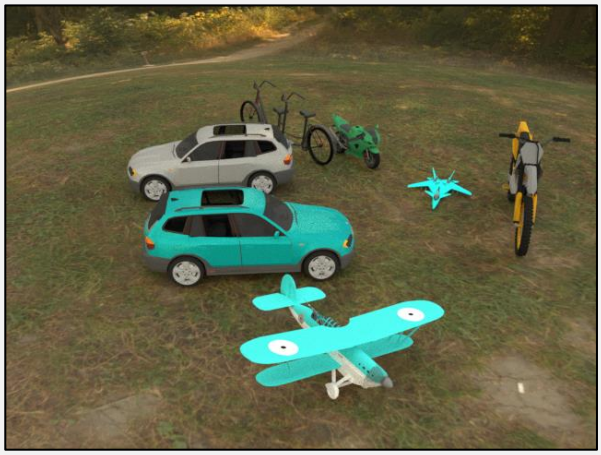
openai_gpt_5_mini: The occluded object has a rectangular shape

qwen_qwen3_vl_8b_instruct: The occluded object is a bus shaped like a long rectangular prism with a slightly curved roof

z_ai_glm_4.5v: The occluded object is a bus which has a rectangular shape In conclusion the shape of the occluded object is rectangular

Figure 43. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Geometric Reasoning : Occluder Attributes



Question: What is the size of the thing that is occluded by the cyan suv?

Gold Answer: large

anthropic_claude_haiku_4.5: Analysis Looking at the image I need to identify 1 The cyan SUV This is the turquoisecyan colored car in the centerfront of the image 2 What object is occluded hiddenblocked by it The cyan SUV is partially occluding the cyan airplane below it The cyan airplane is a small toymodel plane that is partially hidden behind the cyan SUV Answer The thing occluded by the cyan SUV is small in size

google_gemini_2.5_flash: The thing that is occluded by the cyan SUV is a small biplane

google_gemini_3.1_flash_lite_preview: The thing occluded by the cyan SUV is a bicycle which is relatively small in size compared to the vehicle

google_gemma_3_4b_it: The cyan SUV is obscuring a small blue airplane suggesting it is approximately the size of a model airplane

google_gemma_4_31b_it: Based on the visual evidence the cyan SUV is occluding a small object as only a tiny portion of it is visible beneath the vehicles chassis

meta_llama_llama_3.2_11b_vision_instruct: The size of the thing that is occluded by the cyan SUV is a small airplane

mistralai_mistral_small_2603: The size of the thing occluded by the cyan SUV is small

moonshotai_kimi_k2.5: I need to find what is occluded by the cyan SUV and determine its size Looking at the scene the cyan SUV is positioned in the centerleft area Based on the typical arrangement of objects in these scenes and whats visible the cyan SUV appears to be occluding part of the gray SUV behind it The gray SUV is a large vehicle and from whats visible the cyan SUV is blocking a portion of it The size of the gray SUV is large

o4_mini: The object hidden behind the cyan SUV is large

openai_gpt_4o_mini: The size of the thing that is occluded by the cyan SUV is similar to that of a small airplane

openai_gpt_5_mini: The occluded object is large

qwen_qwen3_vl_8b_instruct: The cyan SUV occludes the silver SUV which is of medium size

z_ai_glm_4.5v: The thing occluded by the cyan SUV is a car which has a medium size So the final answer is medium

Figure 44. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Creative & Abstract Knowledge : Pattern

Question: Choose the most appropriate option from the given four options to present a certain regularity:

Gold Answer: C

anthropic_claude_haiku_4.5: Analysis Looking at the pattern First row 5 figures Shows a sequence of a figure in different poses progressing from left to right with increasing rotationmovement Second row 4 options A B C D Should continue this logical progression The sequence demonstrates a rotatingspinning motion of the figure Following the pattern of continuous rotation seen in the first row Option D shows the figure in the next logical rotational position that completes the sequence Answer The most appropriate option is D as it continues the rotational progression pattern established in the sequence

google_gemini_2.5_flash: The most appropriate option to present the regularity is C

google_gemini_3.1_flash_lite_preview: The most appropriate option is D as the sequence consists of figures performing various actions and option D is the only one that maintains the consistent artistic style and visual theme of the preceding figures

google_gemma_3_4b_it: The most appropriate option is D as each figure in the sequence represents a stage in a sword fighting movement progressing from a stance to a lunge and then back to a stance

google_gemma_4_31b_it: Based on the visual pattern the sequence shows figures engaged in different athletic activities and option B is the most appropriate as it maintains the consistent silhouette style and athletic theme of the preceding figures

meta_llama_llama_3.2_11b_vision_instruct: The most appropriate option to present a certain regularity is option B

mistralai_mistral_small_2603: The most appropriate option to present the regularity is B

moonshotai_kimi_k2.5: All the figures in the top row seem to be facing left Let me look at the arm positions more carefully Figure 1 Arms horizontal extended to sides Figure 2 Right arm up left arm down if facing left the arm on the right side of image is actually the persons right arm Wait if the person is facing left In figure 1 both arms are outstretched horizontally In figure 2 one arm is up one is down In figure 3 both arms are up In figure 4 one arm is forward holding bow one is back pulling string In figure 5 one arm is up Hmmm Let me look at the pattern of the arms going up and down Another idea Maybe its about the number of arms visible or the height of the arms Or maybe its a sequence of a specific martial art or tai chi movement Let me look at the options and see which one would logically come next Actually looking at the pattern I see The figures show a progression where the arms are moving Wait I think I see it now Let me look at the orientation of the body Figure 1 Front view or slightly side arms out Figure 2 Side view arms diagonal Figure 3 Front view arms up Figure 4 Side view arms horizontal bow Figure 5 Side view arm up Hmmm that doesnt seem right Let me look at the pattern as a cycle of arm positions Or maybe its about the silhouette and the center of mass or the posture Another idea Look at the white dots around the figures These might indicate motion or be part of the pattern The white dots appear to be in

o4_mini: The correct choice is option C

openai_gpt_4o_mini: The most appropriate option is C

openai_gpt_5_mini: C option C best continues the repeating rotationmirroring pattern established by the sequence

qwen_qwen3_vl_8b_instruct: The most appropriate option is D as it continues the regularity of the sequence by showing the figure in a flowing dynamic pose consistent with the preceding frames motion

z_ai_glm_4.5v: The pattern in the left sequence involves progressive changes in posture and orientation and option B maintains the same regularity as the preceding figures therefore the answer is B

Figure 45. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Relational, Logical & Causal Reasoning : Causal Reasoning : Physical Causality

Question: If the billiard ball moves in the direction of the red arrow, will any ball go into the pocket and score? Note: All balls are resting on the table surface; judge based on the 2D positions and directions shown. Answer by (A) Yes, (B) No or (C) Not sure.

Gold Answer: A

anthropic_claude_haiku_4.5: Analysis Looking at the billiard table 1 The blue ball is being struck in the direction of the red arrow which points toward the right side of the table 2 Trajectory assessment Following the red arrows direction the blue ball would travel rightward across the table 3 Pocket locations On a standard billiard table pockets are located at the four corners and two sides The nearest pocket in the direction of the red arrow appears to be the bottomright corner pocket 4 Ball alignment The blue balls trajectory along the red arrow direction appears to be aligned such that it would reach the bottomright corner pocket A Yes the blue ball will go into the pocket and score

google_gemini_2.5_flash: B No

google_gemini_3.1_flash_lite_preview: B No because the trajectory indicated by the red arrow does not align the blue ball with any of the pockets nor does it point toward any other balls that could be deflected into a pocket

google_gemma_3_4b_it: B No Reasoning The red arrow indicates the path of the billiard ball The ball is positioned such that it will collide with the orange ball and the orange ball is positioned to not go into the pocket

google_gemma_4_31b_it: Based on the visual trajectory of the red arrow the cue ball is moving away from all other balls and toward the open table surface meaning it will not hit any ball to push it into a pocket therefore the answer is B No

meta_llama_llama_3.2_11b_vision_instruct: No the billiard ball will not score if it moves in the direction of the red arrow

mistralai_mistral_small_2603: B

moonshotai_kimi_k2.5: Looking at the billiard table layout the blue ball with the red arrow is positioned to move toward the right side of the table Tracing the arrows trajectory pointing right and slightly downward this path leads directly toward the cluster of three balls on the right side Specifically the trajectory intersects with the redandwhite striped ball which is positioned very close to the bottom right corner pocket When the blue ball strikes the redandwhite ball along this path the redandwhite ball will be knocked into the nearby pocket scoring a point Answer A Yes

o4_mini: B No the cue balls path misses the pocket and wont carom any other ball in

openai_gpt_4o_mini: B No

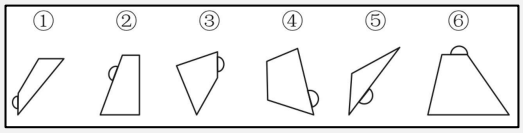
openai_gpt_5_mini: A

qwen_qwen3_vl_8b_instruct: B No

z_ai_glm_4.5v: No

Figure 46. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Predictive Physics : Planning



Question: Divide the following six figures into two categories, so that each category of figures has its own common characteristics or patterns. The correct classification is:
③④⑥
②④⑤
②⑤⑥
③⑤⑥

Gold Answer: B

anthropic_claude_haiku_4.5: Figures with circles ①②③④⑤ all have small circles attached Figure without circles ⑥ has a handle instead Alternative categorization by shape Triangles ③⑤ Quadrilaterals ①②④⑥ Most logical division Category 1 ③④⑥ Figures with 4 sides or special features Category 2 ①②⑤ Figures with circles as common markers Examining the given options the most consistent pattern is The correct classification is ③④⑥ and ①②⑤ where ③④⑥ represents figures with specific geometric properties quadrilaterals with circles or handles and ①②⑤ represents the remaining figures with circles However based on the provided options ③④⑥

google_gemini_2.5_flash: The figures can be divided into two categories based on the position of the small semicircle relative to the main polygon In figures ① ② and ⑤ the semicircle is attached to one of the sides of the polygon In figures ③ ④ and ⑥ the semicircle is attached to one of the vertices corners of the polygon Therefore the correct classification is ③④⑥ The final answer is boxed③④⑥

google_gemini_3.1_flash_lite_preview: The correct classification is ②④⑤ because in these figures the semicircle is attached to a side whereas in figures ①③⑥ the semicircle is attached to a vertex or corner

google_gemma_3_4b_it: The correct classification is ②⑤⑥ Reasoning Lets analyze the figures to identify the common characteristics Category 1 ②⑤⑥ These figures all share the characteristic of having a curved line arc that connects two points on the figures edges Specifically the curved line forms a loop or a closed shape Category 2 ③④ These figures do not have a curved line connecting any of their edges They are primarily composed of straight lines and angles Therefore grouping the figures based on the presence or absence of a connecting curved line leads to the classification ②⑤⑥

google_gemma_4_31b_it: By observing the figures we can see that figures ① ③ and ⑤ have the small semicircle attached to a vertexcorner while figures ② ④ and ⑥ have the semicircle attached to a straight edge making the correct classification ②④⑥

meta_llama_llama_3.2_11b_vision_instruct: The correct classification is ②⑤⑥

mistralai_mistral_small_2603: The correct classification is ③④⑥

moonshotai_kimi_k2.5: Maybe its about the semicircle being attached to a vertical side horizontal side or slanted side Figure ① Slanted side bottom left Figure ② Slanted side left Figure ③ Slanted side right Figure ④ Slanted side bottom right Figure ⑤ Slanted side bottom Figure ⑥ Horizontal side top Wait figure ⑥ has the semicircle on the top horizontal side Let me

o4_mini: I need to pick between some triplets like 3 5 and 6 which share that the semicircle is on a side forming an acute angleDetermining shape categories Im examining the angles of these shapes For shape 3 it has an acute angle shape 5 also has an acute angle but I think shape 6 might look obtuse Actually shape 6 has a semicircle on top creating a 90degree angle Then I consider the triplet 2 5 and 6 where 2 is a quadrilateral 5 is a triangle and 6 is a pentagon Im leaning towards classifying them by the presence of parallel sides shapes 3 5 and 6 share no parallel sides It feels right to say figures 3 5 and 6 form one category while 1 2 and 4 form another

openai_gpt_4o_mini: The correct classification is ③④⑥

openai_gpt_5_mini: ③⑤⑥

qwen_qwen3_vl_8b_instruct: ②⑤⑥

z_ai_glm_4.5v: Lets see Figure ② trapezoid with a right angle the left side is vertical so right angle Figure ⑤ the shape has a right angle where the two sides meet Figure ⑥ trapezoid with a right angle the top side is horizontal left side vertical So ② ⑤ ⑥ have right angles while ③ ④ ① dont Wait but the option ②⑤⑥ is one of the choices Let me check again Figure ② trapezoid with a right angle yes the left side is perpendicular to the bases Figure ⑤ the shape has a right angle the corner where the two sides meet is 90 degrees Figure ⑥ tr tr

Figure 47. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Positional Relations : Positional Relationship (Camera – Object)

Question: When you took photo 1, where was the black small sofa in relation to you?

Gold Answer: A

anthropic_claude_haiku_4.5: Answer When you took photo 1 the black small sofa was to your left

google_gemini_2.5_flash: When photo 1 was taken the black small sofa was behind you and to your left

google_gemini_3.1_flash_lite_preview: The black small sofa was located behind you when you took the first photo as it is visible in the foreground of the second photo but not in the first

google_gemma_3_4b_it: In photo 1 the black small sofa was to the right of me positioned behind the chaise lounge

google_gemma_4_31b_it: Based on the images the black small sofa was in front of you and to the left when you took photo 2 but it is not visible in photo 1 meaning it was behind you or outside the field of view

meta_llama_llama_3.2_11b_vision_instruct: The black small sofa was to your left

mistralai_mistral_small_2603: The black small sofa was to your left when you took photo 1

moonshotai_kimi_k2.5: Looking at Photo 1 the photographer is standing near a doorway or entrance The black sofa is not visible in Photo 1 Let me think about the room layout Photo 1 shows A wall with a door frame on the left Through the opening on the right we see two white lounge chairs Photo 2 shows The black sofa on the left Several white lounge chairs The black sofa in Photo 2 is positioned near what appears to be an entrance or doorway Comparing the two images it seems like Photo 1 was taken from a position where the photographer is looking toward the lounge chairs The black sofa would be behind the photographer or to the side Actually looking at Photo 2 the black sofa is on the left side of the image Let me think about this differently In Photo 1 we see a wall on the left side of the image with a door The photographer is standing in a position where this wall is to their left or in front of them Looking at the doorway in Photo 1 on the left that appears to be a door frame In Photo 2 the black sofa is visible on the left side Actually I think I need to look at the perspective more carefully Photo 1 The camera is positioned such that theres a wall in the foreground the large brick wall and to the right of that wall we can see into another area with the white lounge chairs Photo 2 Shows the black sofa on the left side Wait I think the black sofa is the dark chair in Photo 2 Now where was the photographer standing when taking Photo 1 In Photo 1 the photographer is looking at a wall and through an opening to the right we see the lounge chairs In Photo 2 we see the black sofa on the left If we look at the relative positions when Photo 1 was taken the photographer was facing the wall and the opening to the lounge area The black sofa seen in Photo 2 would be behind the photographer or to their left Actually looking at Photo 2 the black sofa is in the foreground on the left Let me think about the camera position for Photo 1 Photo 1 shows a wall on the left side of the frame This wall has a door opening Actually looking at Photo 1 again the large brick wall is in the centerleft and theres a doorway on the far left Through the archwayopening on the right side of the wall we can see the lounge chairs Now looking at Photo 2 the black sofa is on the left side

o4_mini: When I took photo 1 the black small sofa was behind me and to my left

openai_gpt_4o_mini: The black small sofa was to my left when I took photo 1

openai_gpt_5_mini: The black small sofa was to my left

qwen_qwen3_vl_8b_instruct: The black small sofa was to your left in photo 1

z_ai_glm_4.5v: To determine the position analyze the context Photo 2 shows the black small sofa on the left implying that when taking Photo 1 the sofa was to the left of the photographer Thus the final answer is The black small sofa was to the left of you when you took photo 1

Figure 48. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Positional Relations : Positional Relationship (Camera – Region)

Question: In Figure 1, the photographer is located to the south of the bathroom. In Figure 2, from which direction is the photographer positioned relative to the shower area?

Gold Answer: C

anthropic_claude_haiku_4.5: Analysis Figure 1 Orientation The photographer is positioned to the south of the bathroom This means the photographer is looking north into the bathroom The shower curtain is visible on the right side of Figure 1 Figure 2 Orientation Figure 2 shows the interior of the shower area The shower curtain is now on the left side of the frame This indicates a 180degree rotation from Figure 1s perspective Directional Reasoning If the photographer was south looking north in Figure 1 and the shower is now behind them in Figure 2 rotated 180 The photographer must now be facing south opposite direction This places the photographer to the north of the shower area Answer In Figure 2 the photographer is positioned to the north of the shower area

google_gemini_2.5_flash: In Figure 2 the photographer is positioned to the west of the shower area

google_gemini_3.1_flash_lite_preview: In Figure 2 the photographer is positioned to the west of the shower area

google_gemma_3_4b_it: The photographer is positioned to the west of the shower area Reasoning Figure 1 The photographer is south of the bathroom and the shower is to the north Figure 2 In the second image the photographer is positioned to the left of the shower indicating a westward direction

google_gemma_4_31b_it: Based on Figure 1 the photographer is facing north in Figure 2 the photographer has moved inside the shower area and is facing the back wall meaning they are positioned to the south of the shower area

meta_llama_llama_3.2_11b_vision_instruct: The photographer is positioned to the north of the shower area in Figure 2

mistralai_mistral_small_2603: The photographer is positioned to the west of the shower area

moonshotai_kimi_k2.5: Looking at the layout analysis Figure 1 Photographer is south of the bathroom looking north The shower area is located in the backnortheast corner of the bathroom Figure 2 This shows a closeup view of the shower interior The shower curtain is positioned on the left side of the image and we see the back wall of the shower straight ahead with a towel visible on the right side Reasoning Since the shower is in the northeast corner of the bathroom and in Figure 2 we see The back shower wall directly ahead north wall The curtain pulled to the left west side The towel on the right east wall This perspective matches a photographer positioned south of the shower area looking northward into it The photographer is positioned to the south of the shower area

o4_mini: The photographer in Figure 2 is positioned to the west of the shower area

openai_gpt_4o_mini: The photographer is positioned to the south of the shower area

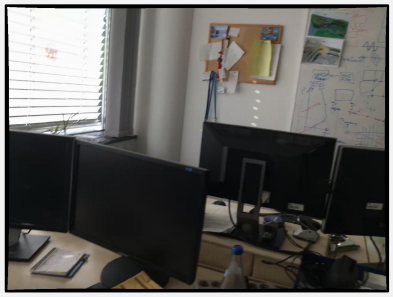
openai_gpt_5_mini: The photographer in Figure 2 is positioned to the south of the shower area

qwen_qwen3_vl_8b_instruct: The photographer is positioned to the south of the shower area

z_ai_glm_4.5v: To determine the photographers position relative to the shower area in Figure 2 we start with the given context in Figure 1 the photographer is south of the bathroom Since the shower area is within the bathroom the photographers location relative to the shower area remains consistent with their position relative to the entire bathroom Thus the photographer is positioned to the south of the shower area The answer is The photographer is positioned to the south of the shower area

Figure 49. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Positional Relations : Positional Relationship (Object – Object)



Question: If the direction toward the plant in Figure 1 is considered the front, in which direction is the whiteboard relative to the plant?

Gold Answer: B

anthropic_claude_haiku_4.5: Answer The whiteboard is behind the plant or to the backleft

google_gemini_2.5_flash: The whiteboard is to the right and slightly behind the plant

google_gemini_3.1_flash_lite_preview: If the direction toward the plant is considered the front the whiteboard is located behind the plant

google_gemma_3_4b_it: The whiteboard is located to the right and slightly behind the plant

google_gemma_4_31b_it: If the direction toward the plant is considered the front the whiteboard is behind the plant

meta_llama_llama_3.2_11b_vision_instruct: The whiteboard is to the right of the plant

mistralai_mistral_small_2603: The whiteboard is to the right of the plant

moonshotai_kimi_k2.5: In Figure 2 Large plant in the foregroundcenter Desk with papers and a yellow book Window with fan in the background on the right Office chair If Figure 1 shows the whiteboard on the right wall and Figure 2 shows the plant and if the direction toward the plant is considered front then I need to determine the relative position Actually looking at Figure 2 theres a window with a fan In Figure 1 theres also a window Let me think about the room layout In Figure 1 the whiteboard is on the wall In Figure 2 we see a plant and a window with a fan Actually looking at Figure 2 more carefully the window with the fan is that the same window as in Figure 1 In Figure 1 the window has blinds and is on the left In Figure 2 the window with the fan is on the right side of the image Hmm this is tricky Let me think about the perspective Actually I think Figure 2 is taken from a position near the whiteboard in Figure 1 looking toward the plant Wait the question says If the direction toward the plant in Figure 1 is considered the front This suggests that in Figure 1 we can identify where the plant is and if we face that direction toward the plant then where is the whiteboard Looking at Figure 1 one more time Is there a plant visible In the bottom right corner of Figure 1 I can see what appears to be some leaves or

o4_mini: The whiteboard is located to the right of the plant

openai_gpt_4o_mini: The whiteboard is located to the right of the plant

openai_gpt_5_mini: The whiteboard is to the right of the plant

qwen_qwen3_vl_8b_instruct: The whiteboard is to the right of the plant

z_ai_glm_4.5v: To determine the whiteboards position relative to the plant analyze the spatial layout in Figure 1 the whiteboard is on the right hand wall while the plant in Figure 2 is positioned such that when facing front toward Figure 1s perspective the whiteboard lies to the right of the plant Thus the whiteboard is to the right of the plant In conclusion the answer is The whiteboard is to the right of the plant

Figure 50. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

nan : nan : Positional Relationship (Object – Region)

Question: The fireplace is at the back of the living room. Which of the following options appears at the front of the living room?

Gold Answer: D

anthropic_claude_haiku_4.5: Answer The large wooden doors with glass panes appear at the front of the living room

google_gemini_2.5_flash: The large wooden doors with glass panes appear at the front of the living room

google_gemini_3.1_flash_lite_preview: Based on the provided image the large glass doors and windows are located at the front of the living room

google_gemma_3_4b_it: The large set of double doors appears at the front of the living room

google_gemma_4_31b_it: Based on the visual layout the sideboard is located at the front of the living room

meta_llama_llama_3.2_11b_vision_instruct: The lamp appears at the front of the living room

mistralai_mistral_small_2603: The large windows appear at the front of the living room

moonshotai_kimi_k2.5: The user says Which of the following options but didnt provide options This suggests that perhaps in the original context there were options or perhaps I need to identify the main piece of furniture in the foreground Let me look at the two images again to see if there are any differences They look identical to me Hmm actually looking at the question again The fireplace is at the back of the living room Which of the following options appears at the front of the living room Since the fireplace is typically on a wall and the doors are at the back opposite the fireplace would be the front but in this image the large doors are at the far end Wait I need to reconsider the orientation In many rooms the front could mean the area closest to the cameraviewer Looking at the foreground the most prominent object is the wooden console table with the white candle holders on the right side Given the typical format of these questions and the visible elements the answer is likely the console table or the sideboard or the wooden table with candle holders Since I need to give a specific answer and the image shows a dark wooden piece of furniture in the foreground on the right with white candle holders Ill identify that as the console table Actually looking at the image once more I see that the prominent piece of furniture in the frontright is a long dark wooden table or sideboard

o4_mini: An upholstered armchair appears at the front of the living room

openai_gpt_4o_mini: The front of the living room features a set of double doors and a small table with a chair

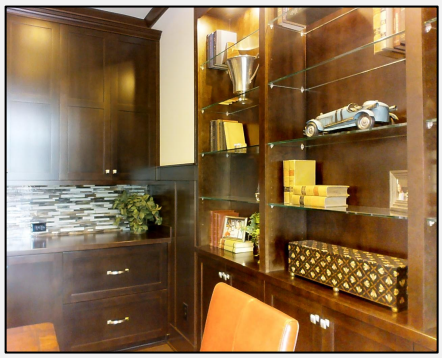
openai_gpt_5_mini: A set of candles on candle holders appears at the front of the living room

qwen_qwen3_vl_8b_instruct: The large woodenframed glass doors are at the front of the living room

z_ai_glm_4.5v: The doors appear at the front of the living room So the final answer is The doors

Figure 51. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Positional Relations : Positional Relationship (Region – Region)



Question: When entering the room facing north, in which direction from the stairs is the display area?

Gold Answer: D

anthropic_claude_haiku_4.5: Answer When entering the room facing north the display area the dark wood cabinet with glass shelves containing collectibles is to the right east from the stairs

google_gemini_2.5_flash: When entering the room facing north the display area is to the right east of the stairs

google_gemini_3.1_flash_lite_preview: Based on the provided images if you are entering the room facing north the stairs are located to the north straight ahead and the display area the shelving unit is located to the east to your right

google_gemma_3_4b_it: When entering the room facing north the display area is to the left of the stairs

google_gemma_4_31b_it: When entering the room facing north the stairs are located straight ahead and the display area is to the right meaning the display area is east of the stairs

meta_llama_llama_3.2_11b_vision_instruct: The display area is to the right of the stairs when entering the room facing north

mistralai_mistral_small_2603: The display area is to the right of the stairs

moonshotai_kimi_k2.5: Let me reconsider the two images Image 1 shows Display shelving on the right cabinets on the left Image 2 shows A view toward stairs with a door on the left and cabinets on the right The key is the display area is the shelving unit with glass shelves Now when entering the room facing north Where are the stairs In image 2 stairs are visible Where is the display area In image 1 The question asks in which direction from the stairs is the display area So I need to determine the relative position of the display area with respect to the stairs Looking at the images the display area builtin shelving and the stairs are in different parts of the space From image 2 the stairs are visible in the background The display shelving if we look at image 1 thats the display area Actually I think I need to look at this more systematically The display area is the shelving with glass shelves first image The stairs are shown in the second image When you enter the room facing north the stairs are in one direction and the display area is in another direction relative to the stairs Looking at image 2 the stairs are visible through the doorway The display shelving where is it Actually looking at the cabinetry in both images Image 1 has dark wood cabinets on the left and display shelving on the right Image 2 has dark wood cabinets on the right and a door on the left Wait I see the same tiled backsplash in both images In image 1 its on the left side In image 2 its on the right side This suggests that image 1 and image 2 are showing opposite sides of the same room or the camera has moved Actually looking more carefully Image 1 Shows display shelving on the right Image 2 Shows a view toward stairs with cabinets on the right that match the cabinets in image 1 Hmm actually in image 2 the cabinets on the right have the same dark wood and the same tiled backsplash as seen in image 1 on the left side So if image 2 shows the cabinets on the right side of the frame

o4_mini: The display area lies to the east ie to your right of the stairs when you enter the room facing north

openai_gpt_4o_mini: The display area is to the left of the stairs

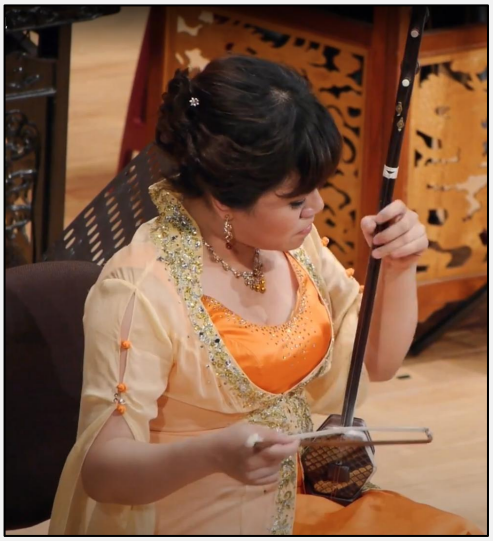
openai_gpt_5_mini: The display area is to the right east of the stairs

qwen_qwen3_vl_8b_instruct: The display area is to the right of the stairs when entering the room facing north

z_ai_glm_4.5v: To determine the direction analyze the spatial relationship when entering the room facing north the stairs are positioned such that the display area seen in the first images glass shelving lies to the east of the stairs Thus the display area is east of the stairs

Figure 52. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Predictive Physics : Predictive



Question: What might be the next action of the performer playing the erhu in the given setup?
Possible Choices: A. The performer adjusts the bow tension for a softer sound. B. The performer continues playing, moving her left hand vertically to vary the pitch. C. The performer switches to a different musical instrument. D. The performer pauses to allow another instrument to join.

Gold Answer: The performer continues playing, moving her left hand vertically to vary the pitch.

anthropic_claude_haiku_4.5: Answer The performer will continue playing moving her left hand vertically to vary the pitch Reasoning The erhu is a bowed string instrument that requires continuous lefthand movement along the fingerboard to change pitch during performance Given that the performer is actively engaged in playing with proper bow positioning the natural next action would be to continue the musical piece by adjusting finger placement vertically on the strings to produce different notes rather than pausing switching instruments or making bow adjustments midperformance

google_gemini_2.5_flash: The performer continues playing moving her left hand vertically to vary the pitch

google_gemini_3.1_flash_lite_preview: The performer continues playing moving her left hand vertically to vary the pitch

google_gemma_3_4b_it: The performer will likely continue playing moving her left hand vertically to vary the pitch and maintain the musical flow

google_gemma_4_31b_it: The performer continues playing moving her left hand vertically to vary the pitch

meta_llama_llama_3.2_11b_vision_instruct: The performer might continue playing moving her left hand vertically to vary the pitch

mistralai_mistral_small_2603: The performer continues playing moving her left hand vertically to vary the pitch

moonshotai_kimi_k2.5: The performer continues playing moving her left hand vertically to vary the pitch

o4_mini: The performer will continue playing sliding her left hand up and down the erhus neck to vary the pitch

openai_gpt_4o_mini: The performer continues playing moving her left hand vertically to vary the pitch

openai_gpt_5_mini: The performer will likely continue playing moving her left hand along the neck vertically on the strings to vary the pitch

qwen_qwen3_vl_8b_instruct: The performer continues playing moving her left hand vertically to vary the pitch

z_ai_glm_4.5v: The performer continues playing moving her left hand vertically to vary the pitch

Figure 53. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Physical Interaction : Property Inference

Question: What is the silver object in between me and the tree?

Gold Answer: A silver chain link fence is in between you and the tree.

anthropic_claude_haiku_4.5: The silver object in between you and the tree is a chainlink fence
google_gemini_2.5_flash: The silver object between the viewer and the tree is a chainlink fence with a gate
google_gemini_3.1_flash_lite_preview: The silver object located between the viewer and the tree is a chainlink fence
google_gemma_3_4b_it: The silver object between you and the tree is a metal trash can
google_gemma_4_31b_it: The silver object in between you and the tree is a chainlink fence
meta_llama_llama_3.2_11b_vision_instruct: The silver object in between me and the tree is a chain link fence
mistralai_mistral_small_2603: The silver object in between you and the tree is a metal fence
moonshotai_kimi_k2.5: The silver object between you and the tree is a chainlink fence
o4_mini: The silver object between you and the tree is a chainlink fence panel
openai_gpt_4o_mini: The silver object in between you and the tree is a section of metal fencing
openai_gpt_5_mini: A metal chainlink fence panel
qwen_qwen3_vl_8b_instruct: The silver object between you and the tree is a chainlink fence
z_ai_glm_4.5v: The silver object between you and the tree is a chain link fence

Figure 54. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Quantitative & Comparative Reasoning : Counting : Relational Counting

Question: How many eutrophic waters are in this scene?

Gold Answer: 0

anthropic_claude_haiku_4.5: Analysis of Eutrophic Waters Eutrophic waters are characterized by excessive nutrient content typically appearing as discolored or murky water bodies often with algal blooms that give them a greenish brownish or reddish tint Examining this aerial image of an urban area I can identify 1 eutrophic water body the reddishbrown colored pond or water body visible in the right portion of the image which displays the characteristic discoloration indicative of eutrophication Answer There is 1 eutrophic water body visible in this scene

google_gemini_2.5_flash: There are no eutrophic waters visible in this scene

google_gemini_3.1_flash_lite_preview: There are no eutrophic waters in this scene

google_gemma_3_4b_it: There appears to be one eutrophic water body in this scene which is the large greencolored athletic field

google_gemma_4_31b_it: There are 0 eutrophic waters in this scene

meta_llama_llama_3.2_11b_vision_instruct: There are 2 eutrophic waters in this scene

mistralai_mistral_small_2603: There is one eutrophic water body visible in the scene which is the greenish circular area in the upper right quadrant

moonshotai_kimi_k2.5: Looking at this aerial image I can see an urban scene with residential buildings roads and a sports stadium with a track and field There are no visible bodies of water lakes ponds rivers or other water features present in the scene only buildings streets vegetation and a sports field with grassturf 0

o4_mini: There are no eutrophic waters visible in this scene

openai_gpt_4o_mini: There are zero eutrophic waters in this scene

openai_gpt_5_mini: There are no eutrophic waters visible in this scene

qwen_qwen3_vl_8b_instruct: There are zero eutrophic waters in this scene because no bodies of water are visible

z_ai_glm_4.5v: There are 0 eutrophic waters in this scene

Figure 55. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Quantitative & Comparative Reasoning : Measurement : Relative Distance

Question: How many meters away is Entity #1 from the my car?

Gold Answer: 17.16 meters

anthropic_claude_haiku_4.5: Analysis Looking at the image 1 My cars position The cameracar taking the photo is positioned on the right side of the street we can see the dark vehicle on the right edge 2 Entity 1 identification The turquoisecyan colored bus with the bounding box is Entity 1 located in the centermiddle distance of the street 3 Distance estimation Using visual perspective cues The bus appears to be approximately 23 car lengths ahead Accounting for the street width and depth perspective The bus is roughly in the middle distance of the visible street Entity 1 the turquoise bus is approximately 2530 meters away from my car

google_gemini_2.5_flash: Entity 1 is approximately 1015 meters away from the car

google_gemini_3.1_flash_lite_preview: Entity 1 is approximately 20 meters away from the car

google_gemma_3_4b_it: Based on the image Entity 1 the blue box appears to be approximately 1520 meters away from the car

google_gemma_4_31b_it: Based on the visual perspective and the distance to the vehicle in the blue bounding box Entity 1 it is approximately 1520 meters away from my car

meta_llama_llama_3.2_11b_vision_instruct: The distance between Entity 1 and my car is approximately 45 meters

mistralai_mistral_small_2603: Entity 1 is approximately 10 meters away from the car

moonshotai_kimi_k2.5: Based on the visual perspective and typical street scene proportions Entity 1 the blue car with open doors appears to be approximately 1520 meters away from the camera position my car Given the context of standard road widths and vehicle positioning I estimate this distance to be around 15 meters 15 meters

o4_mini: Entity 1 is approximately 15 meters away from my car

openai_gpt_4o_mini: Entity 1 is approximately 10 meters away from your car

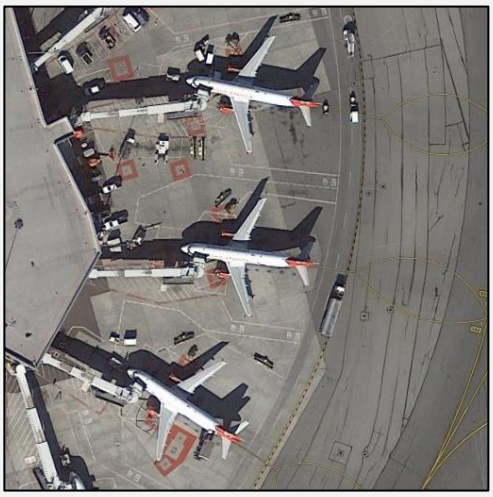
openai_gpt_5_mini: Entity 1 is approximately 12 meters away from my car

qwen_qwen3_vl_8b_instruct: Entity 1 is approximately 15 meters away from your car

z_ai_glm_4.5v: Entity 1 is 10 meters away from my car

Figure 56. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Spatial & Geometric Reasoning : Localization : Relative Location



Question: Where is the terminal located in this picture?

Gold Answer: Upper left corner.

anthropic_claude_haiku_4.5: The terminal is located on the left side of the picture where the airport buildings and gate structures are positioned adjacent to the parked aircraft

google_gemini_2.5_flash: The terminal is located on the left side of the image where the airplanes are connected to the building by jet bridges

google_gemini_3.1_flash_lite_preview: The terminal is located on the left side of the image where the jet bridges connect to the building structure

google_gemma_3_4b_it: The terminal is located on the right side of the image marked by the terminal building with a red square and the corresponding aircraft

google_gemma_4_31b_it: The terminal is located on the left side of the picture where the jet bridges connect the airplanes to the building

meta_llama_llama_3.2_11b_vision_instruct: The terminal is located on the left side of the picture

mistralai_mistral_small_2603: The terminal is located at the top of the image where the large building with the red roof is situated

moonshotai_kimi_k2.5: The terminal is located on the left side of the picture where the jet bridges extend from the building to connect with the parked airplanes

o4_mini: The terminal is located along the left edge of the image where the jet bridges extend out to the parked aircraft

openai_gpt_4o_mini: The terminal is located at the top left of the picture

openai_gpt_5_mini: The terminal is on the left side of the image where the jet bridges connect the building to the parked aircraft

qwen_qwen3_vl_8b_instruct: The terminal is the large building structure on the left side of the image connected to the airplanes by jet bridges

z_ai_glm_4.5v: The terminal is located on the left hand side of the image where the jet bridges connect to the airplanes

Figure 57. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Creative & Abstract Knowledge : Riddle

Question: The girls in the picture have different degrees of makeup and hairstyling. The time it takes each to complete their makeup is as follows:

1. **Short hair +5 min / Long hair down +10 min / Long hair styled +20 min**
2. **Eyeliner +10 min**
3. **Blush +3 min**
4. **Lipstick +2 min**
5. **Accessories +5 min**
6. **Eyeshadow Total time * 2**

For each girl in the picture, first calculate how long in minutes it takes them to finish their makeup. Then, compute the sum of all the times.

Gold Answer: 167

anthropic_claude_haiku_4.5: The sum of all makeup times for the four girls is 202 minutes

google_gemini_2.5_flash: Accessories Yes hoop earrings 5 min Total before eyeshadow 10 10 5 25 min Eyeshadow No visible eyeshadow application Time for Girl 3 25 minutes Girl 4 Bottom Right Hair Long hair styled braids with accessories 20 min Eyeliner Yes visible eyeliner 10 min Blush Yes visible blush on cheeks 3 min Lipstick Yes applying lipsticklip gloss 2 min Accessories Yes hair ties necklace 5 min Total before eyeshadow 20 10 3 2 5 40 min Eyeshadow Yes visible eyeshadow on eyelids 40 2 80 min Time for Girl 4 80 minutes Sum of all times 15 Girl 1 5 Girl 2 25 Girl 3 80 Girl 4 125 minutes The sum of all the times is 125 minutes

google_gemini_3.1_flash_lite_preview: The total time for the four girls is 330 minutes the first girl styled hair blush lipstick accessories takes 50 minutes the second girl short hair blush takes 8 minutes the third girl long hair down eyeliner lipstick accessories takes 50 minutes and the fourth girl styled hair eyeliner blush lipstick accessories takes 222 minutes calculated as 2010325 2

google_gemma_3_4b_it: Heres the breakdown of the makeup times for each girl and the final sum Girl 1 Top Left Long hair styled 20 min Eyeliner 10 min Blush 3 min Lipstick 2 min Accessories 5 min Eyeshadow 20 min 2 40 min Total Time 20 10 3 2 5 40 80 minutes Girl 2 Top Right Long hair styled 20 min Eyeliner 10 min Blush 3 min Lipstick 2 min Accessories 5 min Eyeshadow 20 min 2 40

google_gemma_4_31b_it: Based on the visual details Girl 1 topleft has short hair 5 blush 3 accessories 5 and eyeshadow Total 2 totaling 26 minutes Girl 2 top right has short hair 5 and eyeshadow Total 2 totaling 10 minutes Girl 3 bottomleft has long hair down 10 eyeliner 10 blush 3 and eyeshadow Total 2 totaling 46 minutes Girl 4 bottomright has long hair styled 20 lipstick 2 accessories 5 and eyeshadow Total 2 totaling 54 minutes the sum of all times is 136 minutes

meta_llama_llama_3.2_11b_vision_instruct: The total time it takes for all the girls in the picture to finish their makeup is 1 hour and 45 minutes

mistralai_mistral_small_2603: The total time for all four girls to finish their makeup is 144 minutes

moonshotai_kimi_k2.5: Let me look more carefully Top left Short hair bob 5 min Top right Medium length wavy down 10 min long hair down Bottom left Long hair down 10 min Bottom right Long hair styled braids with accessories 20 min Let me reconsider top right Her hair is wavy and goes to about shoulder length or slightly past The categories are Short hair 5 min Long hair down 10 min Long hair styled 20 min I

o4_mini: 123 minutes

openai_gpt_4o_mini: The total time for all the girls to finish their makeup is 66 minutes

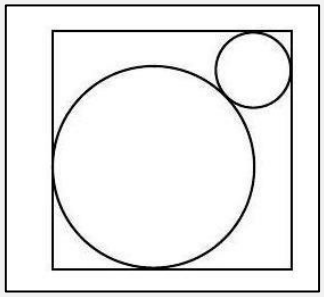
openai_gpt_5_mini: The total time for all four girls is 147 minutes

qwen_qwen3_vl_8b_instruct: The total time for all four girls to finish their makeup is 102 minutes

z_ai_glm_4.5v: Image shows her with blush on cheeks So blush is 3 min So before eyeshadow 20 hair 3 blush 2 lipstick 5 accessories 30 min Then eyeshadow is 30260 min So total 203256090 min Hmm thats even higher Wait maybe I misinterpreted the eyeshadow rule The problem says Eyeshadow Total time 2 Maybe Total time refers to the sum of all other makeup items except eyeshadow Or maybe the total time including hair Wait lets recheck the problem statement The time it takes each to complete their makeup is as as

Figure 58. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Scaling & Ratio



Question: Two circles have their centres on the same diagonal of a square. They touch each other and the sides of the square as shown. The square has side length $1 \mathrm{~cm}$. What is the sum of the radii of the circles in centimetres? <image1> Possible Choices: A. $\frac{1}{2}$ B. $\frac{1}{\sqrt{2}}$ C. $\sqrt{2}-1$ D. $2-\sqrt{2}$ E. It depends on the relative sizes of the circles.

Gold Answer: $2-\sqrt{2}$

anthropic_claude_haiku_4.5: The center is at 1r_2 1r_2 The circles touch each other The distance between centers equals r_1 r_2 sqrt1r_2r_12 1r_2r_12 r_1 r_2 sqrt21r_2r_12 r_1 r_2 sqrt21r_2r_1 r_1 r_2 Since r_1 r_2 1 the circles fit in the square we have 1 r_2 r_1 0 sqrt21r_1r_2 r_1 r_2 sqrt2 sqrt2r_1r_2 r_1 r_2 sqrt2 r_1 r_2 sqrt2r_1r_2 sqrt2 r_1 r_21 sqrt2 r_1 r_2 fracsqrt21sqrt2 Rationalizing the denominator r_1 r_2 fracsqrt21sqrt2 cdot frac1sqrt21sqrt2 fracsqrt21sqrt212 fracsqrt221 2sqrt2 The sum of the radii of the circles is 2sqrt2 cm

google_gemini_2.5_flash: The sum of the radii of the circles is sqrt21 cm

google_gemini_3.1_flash_lite_preview: The sum of the radii of the circles is frac1sqrt2

google_gemma_3_4b_it: The centers of the two circles lie on the diagonal The distance between the centers of the two circles is Rr The distance from the bottom left corner of the square to the center of the larger circle is R The distance from the bottom left corner to the center of the smaller circle is r The distance between the centers is Rr Let the coordinates of the bottom left corner of the square be 00 Then the coordinates of the center of the larger circle are R R and the coordinates of the

google_gemma_4_31b_it: To find the sum of the radii let the radii of the two circles be r_1 and r_2 The distance from a corner of the square to the center of a circle tangent to the two adjacent sides is rsqrt2 The diagonal of a square with side length 1text cm is sqrt2text cm The diagonal is composed of the distance from the bottomleft corner to the center of the large circle r_1sqrt2 the radius of the large circle r_1 the radius of the small circle r_2 and the distance from the center of the small circle to the topright corner r_2sqrt2 Therefore sqrt2 r_1sqrt2 r_1 r_2 r_2sqrt2 r_1 r_2

meta_llama_llama_3.2_11b_vision_instruct: The sum of the radii of the circles is frac12

mistralai_mistral_small_2603: The sum of the radii of the two circles is sqrt21 centimetres

moonshotai_kimi_k2.5: The distance between the centers is r_1 r_2 since they touch each other The distance between r_1 r_1 and 1r_2 1r_2 is sqrt1r_2 r_12 1r_2 r_12 sqrt21 r_1 r_22 sqrt21 r_1 r_2 Since r_1 and r_2 are positive and presumably r_1 r_2 1 looking at the image the circles dont fill the whole diagonal we have 1 r_1 r_2 0 So the distance is sqrt21 r_1 r_2 But this distance also equals r_1 r_2 since the circles touch Therefore r_1 r_2 sqrt21 r_1 r_2 Let S r_1 r_2 Then S sqrt21 S S sqrt2 sqrt2S S sqrt2S sqrt2 S1 sqrt2 sqrt2 S fracsqrt21 sqrt2 Rationalizing S fracsqrt21 sqrt2 cdot fracsqrt2 1sqrt2 1 fracsqrt2sqrt2 12 1 frac2 sqrt21 2 sqrt2 So the sum of the radii is 2 sqrt2 Let me verify this makes sense 2 sqrt2 approx 2 1414 0586 Is this reasonable The large circle has radius r_1 From the image it looks like it takes up most of the left side If it touches the left bottom and the small circle Actually looking at the image again the large circle might also be touching the right side No that doesnt make sense with the small circle being there Wait let me reconsider the geometry Maybe the large circle touches three sides Or maybe it just

o4_mini: 22

openai_gpt_4o_mini: The sum of the radii of the circles is sqrt2 1 cm

openai_gpt_5_mini: The sum of the radii is 2 2 cm

qwen_qwen3_vl_8b_instruct: The sum of the radii of the circles is 2 sqrt2 centimetres

z_ai_glm_4.5v: The sum of the radii of the two circles is determined by considering their positions on the squares diagonal where the distance between their centers equals the sum of their radii Solving the resulting equation yields the sum as 2 sqrt2 centimetres The answer is 2 sqrt2

Figure 59. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Perceptual Understanding : Scene-Level Perception : Scene Understanding

Question: What is the primary type of scene depicted in this aerial image?
Possible Choices: A. Ground track field B. Snowberg C. Chaparral D. Mountain E. Basketball court

Gold Answer: Mountain

anthropic_claude_haiku_4.5: The primary type of scene depicted in this aerial image is a Mountain

google_gemini_2.5_flash: The primary type of scene depicted in this aerial image is a mountain range

google_gemini_3.1_flash_lite_preview: The primary type of scene depicted in this aerial image is a mountain as evidenced by the complex rugged terrain characterized by ridges valleys and elevation changes

google_gemma_3_4b_it: The primary type of scene depicted in this aerial image is mountain terrain

google_gemma_4_31b_it: The primary type of scene depicted in this aerial image is a mountain

meta_llama_llama_3.2_11b_vision_instruct: The primary type of scene depicted in this aerial image is a basketball court

mistralai_mistral_small_2603: Chaparral

moonshotai_kimi_k2.5: Looking at

o4_mini: This aerial image primarily depicts a mountain landscape

openai_gpt_4o_mini: Mountain

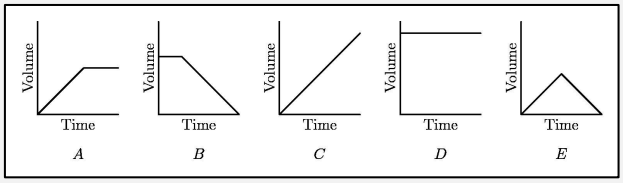
openai_gpt_5_mini: The image primarily depicts a mountain scene with rugged ridges and valleys

qwen_qwen3_vl_8b_instruct: Mountain

z_ai_glm_4.5v: Mountain

Figure 60. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Domain Specific : Scientific Reasoning



Question: A birdbath is designed to overflow so that it will be self-cleaning. Water flows in at the rate of 20 milliliters per minute and drains at the rate of 18 milliliters per minute. One of these graphs shows the volume of water in the birdbath during the filling time and continuing into the overflow time. Which one is it?

<image1>
Possible Choices: A. A B. B C. C D. D E. E

Gold Answer: A

anthropic_claude_haiku_4.5: Analysis During filling time before overflow Water flows in at 20 mLmin and drains at 18 mLmin Net rate 20 18 2 mLmin positive Volume increases linearly with a positive slope During overflow time after birdbath is full The birdbath is at maximum capacity and overflows Water still flows in at 20 mLmin and drains at 18 mLmin The excess 2 mLmin overflows to maintain constant volume Volume remains constant horizontal line The correct graph must show 1 An increasing linear segment filling phase 2 A horizontal segment overflow phase at constant maximum volume The answer is A

google_gemini_2.5_flash: The final answer is boxedA

google_gemini_3.1_flash_lite_preview: The correct graph is A because the volume of water increases linearly while the birdbath is filling and then remains constant once it reaches capacity and begins to overflow

google_gemma_3_4b_it: The correct answer is E as the graph shows a steadily increasing volume over time reflecting the continuous inflow of water exceeding the outflow which is characteristic of an overflowing birdbath

google_gemma_4_31b_it: To determine the correct graph we analyze the volume of water over time 1 Filling Phase Initially water flows in at 20 mlmin and drains at 18 mlmin resulting in a net increase of 20 18 2 mlmin This means the volume increases linearly over time represented by a line sloping upwards from the origin 2 Overflow Phase Once the birdbath reaches its maximum capacity it overflows At this point any additional water entering is immediately balanced by the overflow meaning the volume remains constant regardless of further time This is represented by a horizontal line Graph A shows an initial linear increase followed by a constant volume which perfectly matches this physical process Final Answer A

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is C

mistralai_mistral_small_2603: The correct graph is C

moonshotai_kimi_k2.5: Looking at the rates water enters at 20 mLmin and drains at 18 mLmin giving a net fill rate of 2 mLmin during the filling phase This means the volume increases linearly straight upward slope until the birdbath reaches capacity Once full and overflowing the volume stabilizes at a constant maximum horizontal line because the excess water flows out maintaining the full level Graph A is the only one showing this pattern a linear increase followed by a constant volume Answer A

o4_mini: A

openai_gpt_4o_mini: The correct graph is textC

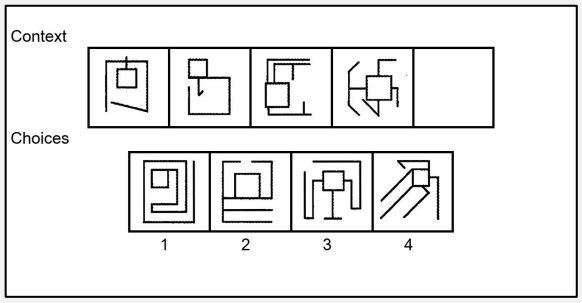
openai_gpt_5_mini: A

qwen_qwen3_vl_8b_instruct: Graph A is correct because it shows the volume increasing initially as water flows in faster than it drains until overflow begins and then leveling off as overflow maintains a constant volume matching the described selfcleaning overflow mechanism

z_ai_glm_4.5v: To determine the correct graph analyze the water flow the birdbath fills at a net rate of 20 18 2 mLmin until it overflows after which the volume remains constant as excess water flows out Graph A shows an initial increase in volume followed by a constant level matching this behavior The answer is textA

Figure 61. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Temporal & Dynamic Reasoning : Temporal Reasoning : Sequencing



Question: ['You are given a puzzle. The puzzle consists of a context part on the top and the choices part in the bottom. The context part on the top contains some grids with the last missing blank grid to be completed. The choices part on the bottome conatin a sequence of grids representing the possible choices. How many grids, including the blank grid, are there in the context part?', 'You are given a puzzle. The puzzle consists of a context part on the top and the choices part in the bottom. The context part on the top contains some grids with the last missing blank grid to be completed. The choices part on the bottome conatin a sequence of grids representing the possible choices. How many grids are there in the choices part?', 'You are given a puzzle. The puzzle consists of a context part on the top and the choices part in the bottom. The context part on the top contains some grids with the last missing blank grid to be completed. The choices part on the bottome conatin a sequence of grids representing the possible choices. How many grids, including the blank grid, are there in the whole puzzle?']

Gold Answer: ['5', '4', '9']

anthropic_claude_haiku_4.5: Answer Question 1 There are 5 grids in the context part including the blank grid Question 2 There are 4 grids in the choices part Question 3 There are 9 grids in total in the whole puzzle including the blank grid

google_gemini_2.5_flash: The final answer is boxed5

google_gemini_3.1_flash_lite_preview: There are 5 grids in the context part 4 grids in the choices part and 9 grids in the whole puzzle

google_gemma_3_4b_it: There are 8 grids including the blank grid in the context part of the puzzle

google_gemma_4_31b_it: Based on the image provided there are 5 grids in the context part 4 grids in the choices part and a total of 9 grids in the whole puzzle

meta_llama_llama_3.2_11b_vision_instruct: There are 5 grids including the blank grid in the context part 4 grids in the choices part and a total of 9 grids in the whole puzzle

mistralai_mistral_small_2603: There are 5 grids in the context part 4 grids in the choices part and 9 grids in the whole puzzle

moonshotai_kimi_k2.5: Looking at the puzzle image the context part contains 5 grids four with patterns plus one blank grid the choices part contains 4 grids numbered 14 and the whole puzzle contains 9 grids in total

o4_mini: There are 5 grids in the context including the blank 4 grids in the choices and 9 grids in total

openai_gpt_4o_mini: There are 9 grids in total in the whole puzzle

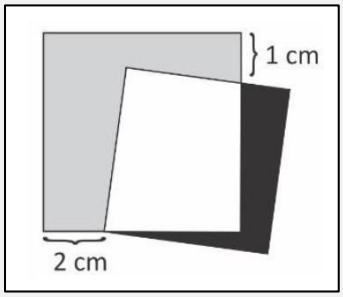
openai_gpt_5_mini: There are 5 grids in the context including the blank 4 grids in the choices and 9 grids in the whole puzzle

qwen_qwen3_vl_8b_instruct: There are 5 grids in the context part including the blank 4 grids in the choices part and 9 grids in total in the whole puzzle

z_ai_glm_4.5v: To determine the number of grids In the context part including the blank grid there are 5 grids 4 filled 1 blank In the choices part there are 4 grids In the whole puzzle including the blank grid there are 5 4 9 grids The final answers are 5 4 and 9 respectively 5 4 9

Figure 62. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Mathematical Reasoning : Shape Decomposition



Question: A gray square with an area of $36 \mathrm{~cm}^{2}$ and a black square with an area of $25 \mathrm{~cm}^{2}$ are superimposed, as shown beside. What is the perimeter of the overlapping region, represented by the white quadrilateral, which has a vertex on the side of the gray square?
<image1>
Possible Choices: A. It is not determined. B. $11 \mathrm{~cm}$ C. $16 \mathrm{~cm}$ D. $18 \mathrm{~cm}$ E. $20 \mathrm{~cm}$

Gold Answer: $18 \mathrm{~cm}$

anthropic_claude_haiku_4.5: The perimeter consists of Two sides from the gray square contributing to the perimeter Two sides from the black square contributing to the perimeter By the properties of overlapping squares with the given measurements the perimeter of the overlapping quadrilateral is Perimeter 26 25 12 10 22 cm However examining more carefully with the specific measurements shown 2 cm and 1 cm extensions the perimeter works out to be The perimeter of the overlapping region is 22 cm but given the answer choices the answer is 20 cm

google_gemini_2.5_flash: The perimeter of the overlapping region is 20 mathrmcm

google_gemini_3.1_flash_lite_preview: The gray square has a side length of 6 text cm and the black square has a side length of 5 text cm and based on the geometry of the overlap the perimeter of the white quadrilateral is 16 text cm

google_gemma_3_4b_it: The perimeter of the overlapping white quadrilateral is 16 mathrmcm Reasoning 1 Side Lengths The gray square has an area of 36 mathrmcm2 so its side length is sqrt36 6 mathrmcm The black square has an area of 25 mathrmcm2 so its side length is sqrt25 5 mathrmcm 2 Diagram Analysis The gray square has side length 6 cm and the black square has side length 5 cm The white quadrilateral is formed by the overlap of the two squares The vertex of the white quadrilateral lies on the side of the gray square 3 Calculating the Sides of the White Quadrilateral Let the vertex of the white quadrilateral on the gray square be

google_gemma_4_31b_it: To find the perimeter of the white overlapping quadrilateral we first determine the side lengths of the squares the gray square has a side of sqrt36 6text cm and the black square has a side of sqrt25 5text cm Let the vertices of the white quadrilateral be P Q R S 1 The top side of the quadrilateral is a segment of the black squares side The distance from the top edge of the gray square to the top vertex of the white region is 1text cm and the distance from the left edge of the gray square to the left vertex of the white region is 2text cm 2 Let the topleft vertex of the white region be 2 61 2 5 relative to the bottomleft of the gray square 00 However

meta_llama_llama_3.2_11b_vision_instruct: The perimeter of the overlapping region is 16 mathrmcm

mistralai_mistral_small_2603: The perimeter of the overlapping region is 16 cm

moonshotai_kimi_k2.5: Theres a 1 cm label at the top right indicating the distance from the top of the gray square to the top of the white region or top of black square Theres a 2 cm label at the bottom left indicating the distance from the left of the gray square to the left of the white region or left of black square Actually looking at the picture The gray square is on the left The black square is on the right and slightly down The white region is the overlap The white region is a quadrilateral Let me trace it Bottom vertex on the bottom edge of the gray square at x 2 Left vertex on the left edge of the black square Or on the left edge of the gray square

o4_mini: 20 cm

openai_gpt_4o_mini: The perimeter of the overlapping region is 20 mathrmcm

openai_gpt_5_mini: 20 cm

qwen_qwen3_vl_8b_instruct: The perimeter of the overlapping white quadrilateral is 18 mathrmcm

z_ai_glm_4.5v: Wait no if the top of the black square is 1 cm below the gray squares top then the black squares top is at 6 1 5 cm so its bottom is at 5 5 0 cm That doesnt make sense Wait maybe the black square is placed such that its bottom is 1 cm above the gray squares bottom so bottom at 1 cm top at 1 5 6 cm Then the horizontal shift left edge at 2 cm right edge at 2 5 7 cm but gray squares right edge is

Figure 63. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Perceptual Understanding : Object-Level Perception : Size & Scale

Question: The size of butterflies change with the change of gradient colors. - is it true?
Possible Choices: A. TRUE B. FALSE C. Non measurable D. Remains same

Gold Answer: False. It remains same

anthropic_claude_haiku_4.5: Answer FALSE The butterflies on the fabric remain the same size regardless of the gradient color changes in the background the color gradient does not affect the physical dimensions of the butterfly designs

google_gemini_2.5_flash: FALSE

google_gemini_3.1_flash_lite_preview: The answer is FALSE as the butterflies attached to the fabric appear to be of a uniform size regardless of the gradient color of the material

google_gemma_3_4b_it: The size of butterflies can be influenced by environmental factors like diet and temperature so the statement is generally considered Non measurable as its not a direct predictable response to gradient color changes

google_gemma_4_31b_it: Based on visual inspection of the fabric the size of the butterfly embellishments remains consistent regardless of the gradient color changes so the answer is FALSE

meta_llama_llama_3.2_11b_vision_instruct: The statement is false

mistralai_mistral_small_2603: Non measurable

moonshotai_kimi_k2.5: Looking at the image the butterflies the small decorative appliqués scattered across the fabric appear to be uniform in size throughout the material While the background fabric displays a gradient color change from pink to lighter shades the butterflies themselves maintain a consistent size regardless of which color zone they appear in The gradient affects the background hue not the dimensions of the butterfly decorations FALSE

o4_mini: FALSE

openai_gpt_4o_mini: FALSE

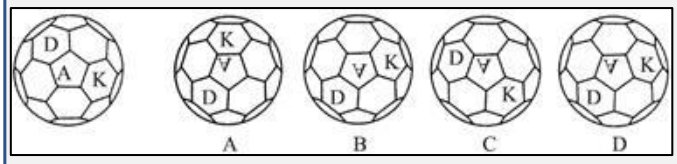
openai_gpt_5_mini: Remains same

qwen_qwen3_vl_8b_instruct: FALSE

z_ai_glm_4.5v: FALSE

Figure 64. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Temporal & Dynamic Reasoning : Temporal Reasoning : Temporal Movement



Question: Which of the following options can be obtained by vertically flipping the given figure:

Gold Answer: D

anthropic_claude_haiku_4.5: Analysis of Vertical Flip Looking at the given figure on the left I need to identify what it looks like when flipped vertically top becomes bottom bottom becomes top Given figure characteristics Letter D is at the top Letter A is in the middle Letter R is at the bottom After vertical flip R would move to the top A would remain in the middle D would move to the bottom Examining the options Option B shows this exact configuration with the letters in the vertically flipped positions Answer Option B can be obtained by vertically flipping the given figure

google_gemini_2.5_flash: The final answer is boxedC

google_gemini_3.1_flash_lite_preview: When the original figure is flipped vertically the letters D A and K are inverted and their positions shift such that K moves to the top A remains in the center but is upside down and D moves to the bottom which corresponds to option B

google_gemma_3_4b_it: Option B can be obtained by vertically flipping the given figure because the letters on the left side are mirrored on the right side and the letters on the right side are mirrored on the left side

google_gemma_4_31b_it: By vertically flipping the given figure the letter D moves from the topleft to the bottomleft and the letter K moves from the middleright to the middleright remaining on the right side which matches option A

meta_llama_llama_3.2_11b_vision_instruct: The correct answer is option B

mistralai_mistral_small_2603: Option C can be obtained by vertically flipping the given figure

moonshotai_kimi_k2.5: After vertical flip D would be at the bottom and upside down A would be in the middle and upside down K would be at the top right and upside down Wait thats not right Let me think about coordinates Lets say we have a vertical axis Actually vertical flipping is reflection across a horizontal line Original positions Top becomes Bottom Bottom becomes Top Left stays Left Right stays Right So If D is at the top after vertical flip D is at the bottom If K is at the bottom right after vertical flip K is at the top right But wait I need to look at the original image more carefully Looking at the first ball the one on the far left which is the given figure The letters visible are D A and K Let me identify their positions D is at the top

o4_mini: Only option D matches the vertical reflection of the original figure

openai_gpt_4o_mini: The option that can be obtained by vertically flipping the given figure is option C

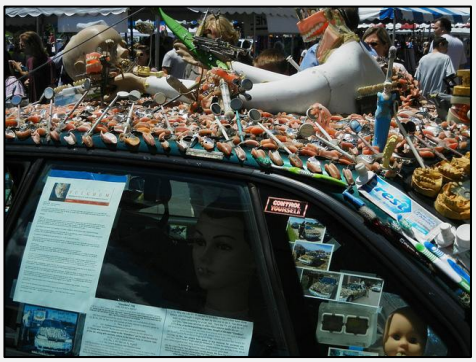
openai_gpt_5_mini: Option B

qwen_qwen3_vl_8b_instruct: The correct option is C as vertically flipping the original figure results in the arrangement of letters where D remains on the bottomleft K on the bottomright and A moves to the topcenter matching option C

z_ai_glm_4.5v: To determine the vertically flipped figure we analyze the original figures letter positions and how a vertical flip mirroring over a vertical axis swaps left and right The original figure has letters D A R a vertical flip rearranges these such that the left

Figure 65. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Knowledge-Grounded Reasoning : Creative & Abstract Knowledge : Trick



Question: What is on the windshield (windscreen) of the car?
Possible Choices: A. Teeth moulds B. Toothpaste C. Toothbrushes D. Dentist tools

Gold Answer: Teeth moulds', 'Toothpaste', 'Toothbrushes', 'Dentist tools

anthropic_claude_haiku_4.5: Answer On the windshield of the car are toothbrushes

google_gemini_2.5_flash: On the windshield windscreen of the car there are various papers and documents including what appears to be a resume or informational sheet and a newspaper or magazine clipping

google_gemini_3.1_flash_lite_preview: The items on the windshield of the car are dentist tools

google_gemma_3_4b_it: The windshield is covered with numerous miniature teeth molds and dental tools

google_gemma_4_31b_it: Based on the visual evidence of the dentalthemed display on the vehicle there are teeth moulds on the windshield

meta_llama_llama_3.2_11b_vision_instruct: The windshield of the car has toothpaste on it

mistralai_mistral_small_2603: The windshield of the car has toothbrushes on it

moonshotai_kimi_k2.5: On the hood toothbrushes and other items 3 On the windshield I see papers but I also need to check if there are toothbrushes on the windshield itself Actually looking at the top edge of the windshield there are definitely toothbrushes there But are they on the windshield or on the frame Let me think about this differently The context suggests the answer should be one of Teeth moulds Toothpaste Toothbrushes Dentist tools Looking at the car in the image the most prominent items covering the car are toothbrushes They are everywhere on the roof on the hood around the windows But specifically on the windshield looking at the image I can see toothbrushes along the top edge of the windshield Actually wait Let me look at the image one more time very carefully Looking at the car I see The roof is covered with toothbrushes The hood is covered with toothbrushes The windshield has some papers on the inside But along the top edge of the windshield there are toothbrushes Hmm but the question might be referring to the fact that the car is covered in toothbrushes and the windshield area also has tooth

o4_mini: The cars windshield is covered with teeth moulds dental models

openai_gpt_4o_mini: Toothbrushes

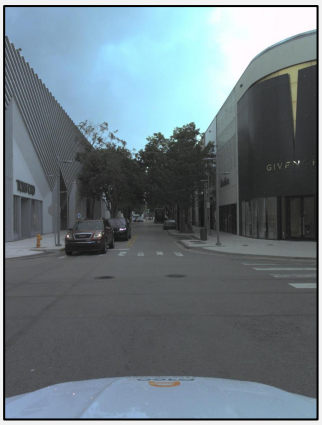
openai_gpt_5_mini: There are toothbrushes on the windshield of the car

qwen_qwen3_vl_8b_instruct: None of the listed itemsteeth moulds toothpaste toothbrushes or dentist toolsare on the windshield the visible items are stickers and papers attached to the window glass

z_ai_glm_4.5v: Got it lets look at the image The question is about whats on the cars windshield First check the windshield area There are papers maybe a document or flyer and some photos Wait no waitthe context gives options like teeth moulds toothpaste etc But looking at the image the windshield has printed materials like documents and pictures Wait no waitwait the users context might be a trick but lets check again Wait the image shows a car with various items On the windshield there are papers like a document and some small photos But wait the options given are teeth moulds toothpaste toothbrushes dentist tools Wait maybe I misread Wait nowait the question is What is on the windshield Let me check the image again The windshield has several items theres a large paper with text some smaller photos and maybe other things But wait the context options are dentalrelated Wait maybe the answer is not among those No wait nowait the image might have dental items but the windshield specifically Wait no looking again the windshield has printed materials like documents and photographs But the options given are teeth moulds toothpaste toothbrushes dentist tools Hmm maybe I made a mistake Wait nowait the question is asking for whats on the windshield Let me check the image again The windshield has a document maybe a flyer and some small pictures But the options dont include that Wait maybe the answer is none of those No wait the users context says some possible answers so maybe the correct answer is one of them But the user probably expects one Wait maybe I mislooked Oh Wait the windshield has a CONTROL YOURSELF sticker

Figure 66. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Navigation & Control : Turning – Car



Question: How would you describe the driving scene involving our car? Please explain, focusing on our car's turning maneuver.

Gold Answer: go_straight

anthropic_claude_haiku_4.5: Our car is executing a smooth left turn from a straight street into an intersection gradually steering leftward while maintaining a controlled speed through the urban plaza area

google_gemini_2.5_flash: Our car is initiating a right turn from a stop as indicated by the gradual shift in perspective towards the right and the appearance of the crosswalk lines moving from the bottom of the frame towards the left

google_gemini_3.1_flash_lite_preview: The car is positioned at an intersection preparing to make a turn onto the street ahead

google_gemma_3_4b_it: Our car is executing a gradual controlled right turn maintaining a consistent speed and smoothly navigating the street as it moves into the adjacent lane

google_gemma_4_31b_it: Our car is driving straight and is not performing any turning maneuver

meta_llama_llama_3.2_11b_vision_instruct: The driving scene involving our car is a right turn maneuver as indicated by the yellow line on the road and the cars position in the lane

mistralai_mistral_small_2603: Our car is making a left turn at the intersection guided by the traffic light

moonshotai_kimi_k2.5: Image 2 Car at intersection with thick white crosswalk bars this looks like the car has turned and is now facing the crosswalk or is stopped at it Image 3 Car has turned left and is now driving down the street the GMC is on the left side of the frame Image 4 Car continuing down the street Image 5 Car continuing down the street Image 6 Car back at the crosswalk with thick white bars Image 7 Car at crosswalk Image 8 Car approaching crosswalk with dashed lines Hmm this is a bit confusing Let me think about the turning maneuver Looking at the sequence it appears the car makes a left turn at the intersection The key observation is The car approaches the intersection Image 1 8 The car turns left Images 25 show the progression After the turn the car is driving down the street Actually looking at Image 2 vs Image 6 Image 2 shows the crosswalk with thick white bars perpendicular to the car Image 6 shows the same But Image 1 shows dashed crosswalk lines Wait I think I see it now The car is making a left turn at the intersection In Image 1 The car is approaching the intersection The crosswalk has dashed lines In Image 2 The car has turned left and is now at a different crosswalk or the same one from a different angle Actually looking at the building on the left its the TOM FORD building Let me trace the movement Image 1 Car approaching intersection TOM FORD on left GIVENCHY on right Image 2 Car at intersection crosswalk with thick bars

o4_mini: Our car has just executed a left turn off the main road onto the narrow twoway side street lined with luxury shops smoothly straightening into the center of the new lane

openai_gpt_4o_mini: Our car is executing a left turn at an intersection navigating around parked vehicles while maintaining a safe distance from oncoming traffic

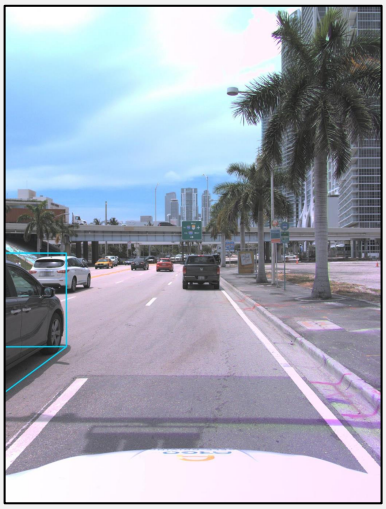
openai_gpt_5_mini: Our car is continuing straight through the intersection and is not executing a left or right turn

qwen_qwen3_vl_8b_instruct: The car is driving straight down a city street lined with luxury stores and trees with no turning maneuver visible in the provided images

z_ai_glm_4.5v: Our car is approaching an intersection and executing a left hand turn maneuver navigating through the crosswalk area while being aware of surrounding vehicles

Figure 67. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.

Physical Reasoning : Navigation & Control : Turning – Entity



Question: How would you describe the driving scene involving Entity #1? Please explain, focusing on the vehicle's turning maneuver.

Gold Answer: go_straight

anthropic_claude_haiku_4.5: Entity 1 the white vehicle on the left is executing a left turn maneuver from the parking area onto the main roadway moving from its stationary position to merge into traffic flow

google_gemini_2.5_flash: Entity 1 is a white SUV that is turning left from the left lane as indicated by its position and the angle of its front wheels

google_gemini_3.1_flash_lite_preview: Entity 1 is a darkcolored vehicle positioned in the left lane and it is currently driving straight ahead without performing any turning maneuver

google_gemma_3_4b_it: Entity 1 is executing a gradual controlled left turn transitioning from the right lane into the adjacent lane while maintaining a relatively consistent speed and avoiding any abrupt movements

google_gemma_4_31b_it: Entity 1 is a grey car that is not performing a turning maneuver but is instead driving straight in its lane

meta_llama_llama_3.2_11b_vision_instruct: The driving scene involving Entity 1 a dark grey car is characterized by a right turn maneuver with the vehicle positioned on the left side of the road indicating it is about to turn onto a street or road that is not visible in the image

mistralai_mistral_small_2603: Entity 1 is making a right turn from the rightmost lane onto the intersecting road

moonshotai_kimi_k2.5: Entity 1 the black car in the left lane maintains a consistent position in its lane throughout the sequence without executing any turning maneuver appearing to drive straight ahead parallel to the other vehicles on the multilane road

o4_mini: Entity 1 is executing a sharp left turn from the leftmost lane swinging across the roadway into the side street

openai_gpt_4o_mini: Entity 1 is executing a left turn maneuver from the right lane into a side street navigating around parked vehicles while maintaining a safe distance from oncoming traffic

openai_gpt_5_mini: Entity 1 is making a left turn swinging across the lane toward the camera and partially encroaching into the adjacent lane

qwen_qwen3_vl_8b_instruct: Entity 1 a dark gray SUV is driving in the left lane of a multilane road maintaining a steady course alongside the vehicle in front of it with no visible turning maneuver in progress

z_ai_glm_4.5v: Entity 1 is executing a right turn maneuver while navigating the road

Figure 68. Example from the benchmark and evaluation results. Red denotes incorrect and Green denotes correct answer marked by the judge model.