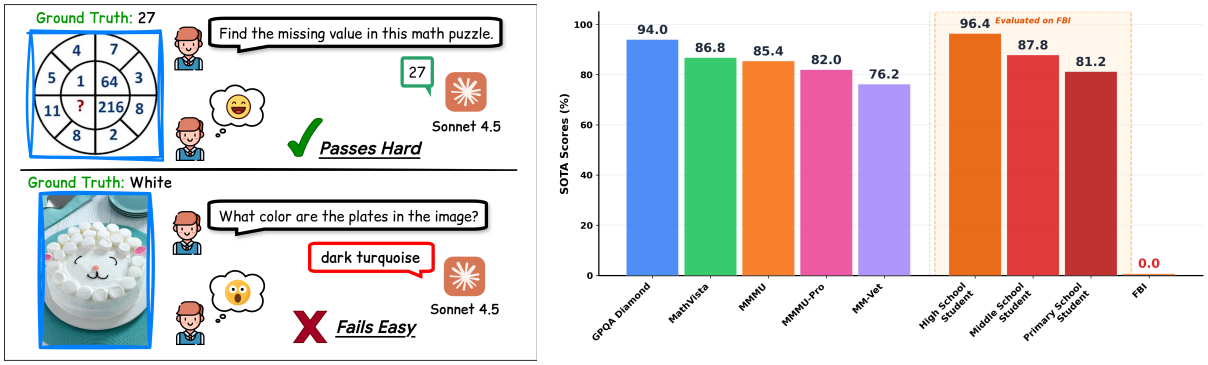
FBI-Bench: Fundamental but Impossible – Are We Overrating VLMs' Capabilities?

Anonymous ACL submission



(a) Motivation

(b) Comparison between FBI-Bench and other widely-used benchmarks

Figure 1: (a) The model solves a complex reasoning task but fails a simple visual question. (b) Despite strong benchmark performance, it achieves 0% on FBI-Bench, revealing a perceptual gap.

Abstract

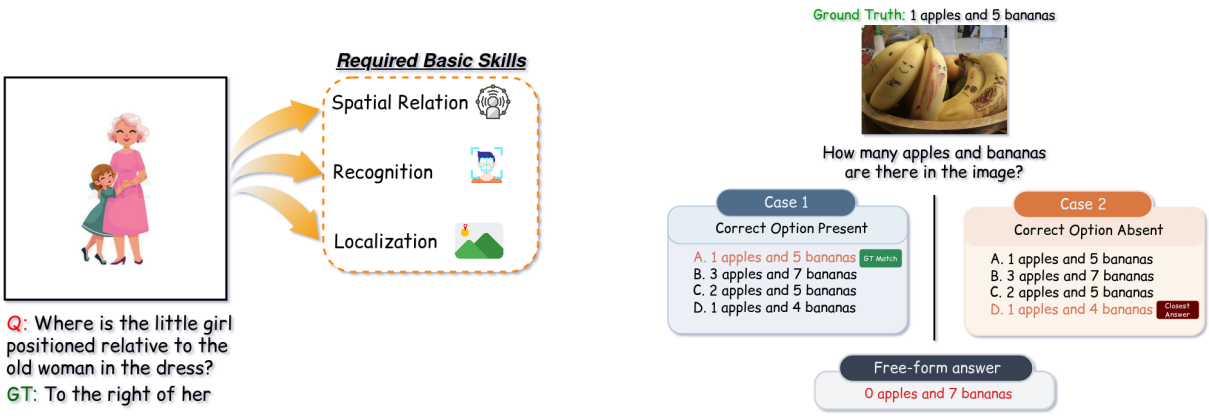
Despite impressive progress on multimodal benchmarks, it remains unclear whether vision–language models (VLMs) possess reliable fundamental perception abilities. Existing benchmarks are often rapidly saturated as models improve, while newer ones increasingly emphasize complex reasoning, which can obscure failures in basic perception. Consequently, fundamental perceptual limitations of VLMs remain difficult to diagnose. To address this issue, we introduce FBI-Bench (Fundamental But Impossible Benchmark), a diagnostic benchmark for evaluating core multimodal skills, including recognition, localization, relation understanding, direction, counting, OCR, and comparison. FBI-Bench contains 500 carefully curated samples with explicit capability annotations, enabling systematic analysis of both individual skills and their compositions. We evaluate a diverse set of state-of-the-art VLMs and find that all models achieve 0.0% accuracy despite the tasks being trivial for humans. To investigate these failures, we conduct controlled experiments examining factors that may influence performance. Our results show that widely used strategies, including failure roots analysis, visual hinting, latent reasoning, post-training, large-scale pretraining, and parameter scaling, provide little improvement in fundamental perception, whereas the design of the vision encoder plays a substantially more critical role. These findings reveal a significant gap between benchmark performance and genuine visual understanding, highlighting the need to strengthen the perceptual foundations of future VLMs.

1 Introduction

"AI isn't as smart as you think, but it could be"

Jeff Dean, 2021

Vision-language models (VLMs) (Liu et al., 2023a; Radford et al., 2021) have significantly rapidly multimodal reasoning nare being adopted in wide range of applications. Given the growing application of VLMs, it is crucial to systematically investigate their limitations. To uncover these, a growing body of benchmarks (Li et al., 2024c,d) has been introduced to evaluate VLMs across diverse tasks and increasingly challenging settings. These benchmarks have played a central



(a) **Fundamental skills** (b) **Evaluation sensitivity**

Figure 2: Basic tasks reveal unstable foundations. (a) A simple question requires only basic perception skills. (b) The model succeeds with answer options but fails in free-form answering.

role in measuring progress and revealing model capabilities. Notably, recent VLMs have reported remarkable performance even on difficult benchmarks (Zhang et al., 2024a; Li et al., 2025). However, strong benchmark performance alone does not necessarily imply a deep or reliable understanding of multimodal inputs. This motivates a simple but important question: **are VLMs actually as strong as we have come to believe?**

Recent headline-grabbing results have further amplified this optimism. Frontier models (Gemini Team, 2025a; Singh et al., 2025) have been reported to achieve gold-medal-level performance on highly demanding reasoning tasks such as IMO (Luong et al., 2025a), fueling speculation that AI systems may be approaching increasingly AGI. But we believe that to achieve AGI, a reliable system, it should not only succeed on exceptionally difficult problems, but also perform consistently on elementary ones. However, as illustrated in **Fig. 1(a)**, highlighting this inconsistency: a frontier model succeeds on a hard puzzle, yet misidentifies an obvious visual attribute. This gap exposes a critical blind spot in current evaluation and motivates the urgent need for a suite that tests whether VLMs truly possess basic visual and multimodal skills. We provide additional motivation for the benchmark design in **Appendix A**.

Existing benchmarks (Yue et al., 2024; Li et al., 2023; Liu et al., 2024c; Yu et al., 2023) increasingly prioritize harder or specialized tasks, under the assumption that greater difficulty leads to more informative assessment. However, this trend can hinder the discovery of fundamental limitations, since strong performance on difficult problems does not guarantee competence in basic skills. Moreover, many benchmarks saturate rapidly (Roberts et al., 2025; Yue et al., 2025), leaving little room for meaningful improvement shortly after their release. We discuss how **FBI-Bench** differs from prior multimodal benchmarks in **Appendix B**.

Building on the above observations, we show that current SOTA VLMs remain far from mastering fundamental skills. Motivated by this gap, we introduce **FBI-Bench** (Fundamental But Impossible), a novel benchmark designed to expose basic limitations that are easily overlooked by existing evaluations. The name reflects our central observation: these skills are fundamental for reliable multimodal intelligence, yet they remain impossible for current frontier models. The contrast in **Fig. 1(b)** is particularly revealing: even primary-school students score above 80% on FBI, whereas frontier models score 0.0, despite excelling on a range of benchmarks. Motivated by recent works of perception as a prerequisite for higher-level cognition (Rao and Ballard, 1999; Barlow, 1990), FBI evaluates 7 fundamental perception skills, including recognition, relation, direction, comparison, OCR, counting, and localization. In addition, following prior work on integrated capability analysis (Yu et al., 2023), we label each sample with the combination of skills it requires, enabling more principled diagnosis of where failures originate **Fig. 2a**.

This issue is further compounded by the evaluation format adopted in many prior benchmarks of basic skills. In particular, existing benchmarks (Fu et al., 2024; Wu et al., 2024) often rely on MCQs, which can yield deceptively strong performance without faithfully measuring whether a model truly

understands the basic skill. Emerging evidence (Yu et al., 2023; Chen et al., 2024b; Wang et al., 2024) suggests that, under multiple-choice settings, models may succeed by selecting approximately correct options, exploiting answer biases, or arriving at the correct choice without genuine understanding, as illustrated in **Fig. 2b**. To mitigate this problem, we construct **FBI-Bench** in a free-form setting and evaluate responses with an LLM-as-a-judge protocol, thereby placing greater emphasis on actual understanding rather than answer-option selection.

Beyond evaluation, we also investigate if basic multimodal skills can be improved using several widely adopted approaches (Zhou et al., 2025b; Abnar et al., 2025; Xie et al., 2024) in current VLM development. Surprisingly, our results show these methods yield little to no meaningful improvement on **FBI-Bench**, despite their effectiveness in other settings (Sun et al., 2023). We further conduct a deeper analysis and find these failures are not merely incidental, but instead point to a persistent weakness in basic capability. Taken together, our findings suggest fundamental multimodal skills remain a major open challenge for current VLMs. We summarize our contributions as follows:

- We show that, despite recent progress, fundamental multimodal skills are not yet reliably acquired by state-of-the-art VLMs, underscoring an important unresolved problem.
- We introduce **FBI-Bench**, a new benchmark for fundamental multimodal skills: FBI is designed to expose basic limitations that are overlooked by existing evaluations, focusing on seven foundational skills, recognition, relation, direction, comparison, OCR, counting, and localization, while remaining challenging enough to avoid rapid saturation.
- We provide fine-grained analysis of failure modes and improvement limits: By annotating each sample with its required skill combinations and testing several widely used improvement strategies, including failure roots analysis, visual hinting, latent reasoning, post-training, large-scale pretraining, and parameter scaling, we show that these approaches yield little meaningful gain in fundamental visual perception. Our analysis further suggests that improvements in vision encoder design play a more decisive role, indicating that fundamental multimodal skills remain a major open challenge for VLMs.

2 Related Works

2.1 Large Vision-language models (LVLMs)

Recent progress in VLMs (Bai et al., 2025a; Zhu et al., 2025) has largely been driven by combining pretrained visual encoders with strong language backbones, followed by multimodal instruction tuning. Early systems(Liu et al., 2023b,a) showed that visual instruction tuning can turn general-purpose language models into capable multimodal assistants, while more recent models further improve grounding and long-context visual understanding. As a result, modern VLMs (OpenAI et al., 2024b; Gemini Team, 2025a) are increasingly viewed as general-purpose multimodal agents rather than task-specific predictors. This rapid progress has enabled impressive performance on a wide range of tasks (Fu et al., 2025b), but it also raises a more basic question: whether these models possess the foundational multimodal skills required for reliable deployment, rather than only excelling on hard cases

2.2 Multimodal Benchmarks

Existing benchmarks (Li et al., 2024c,d) can be broadly grouped into 3 lines. The first line focuses on broad capability coverage (MME(Fu et al., 2025a), MMBench(Liu et al., 2024c), which evaluate models across multiple perception and cognition dimensions at scale). The second line emphasizes challenging multimodal reasoning((Lu et al., 2024; Yue et al., 2025)), and has been central to measuring progress on expert-level tasks. A third line moves toward vision-critical evaluation(Fu et al., 2024; Chen et al., 2024b), which stress perception-heavy problems. Together, these benchmarks have advanced the evaluation of VLMs and made model progress much more measurable. Despite their value, these still leave an important gap for our setting. Many recent benchmarks (Yue et al., 2025, 2024) increasingly prioritize harder tasks, while a large fraction of them rely on MCQ formats for efficient large-scale assessment. Although these design choices are practical, they can make it difficult to determine whether a model truly understands the underlying basic skill, or whether it is benefiting from answer recognition or shortcut strategies. Our work, **FBI-Bench**, targets fundamental, human-easy multimodal skills rather than increasingly difficult tasks, and it evaluates them

Figure 3: Representative examples of the seven foundational capabilities in FBI-Bench. Each panel illustrates a task targeting a specific skill: OCR, counting, comparison, recognition, orientation, localization, and spatial relation.

in a free-form setting with skill-composition annotations, enabling more faithful and fine-grained analysis of where current VLMs fail.

3 FBI-Bench

3.1 Taxonomy

Fundamental Skill Taxonomy: FBI-Bench is built on the premise that *reliable* multimodal intelligence should be consistent on elementary perceptual skills, not only successful on complex demonstrations. Following definitions of core skills in (Zhou et al., 2025a; Li et al., 2024c) and prior views of fundamental human perception (Rao and Ballard, 1999; Barlow, 1990), we organize FBI-Bench around seven foundational capabilities.

Recognition (Rec). Recognition asks **what** is present in the image. It requires identifying the presence, category, state, or type of a visual entity.

Localization (Loc). Localization asks **where** an entity is located. It requires mapping a textual reference to a specific region in the image.

Relation (Rel). Relation understanding asks **how** entities are connected. It requires identifying spatial relations between two or more visual entities.

Direction (Dir). Direction asks **in which** orientation or ordered spatial direction an entity appears, such as left, right, above, below, front, or behind.

Comparison (Com). Comparison asks **which differs more/less** along a visual dimension. It requires comparing attributes or magnitudes.

OCR (Ocr). OCR asks **what the image says**. It requires reading and transcribing visible text in scenes, signs, charts, or other visual contexts.

Counting (Cou). Counting asks **how many** entities satisfy a given visual condition. It requires determining the exact number of objects or regions.

Each item is labeled with the minimal set of abilities required to answer the question. Thus, FBI-Bench supports both single-skill evaluation and ability-integration analysis, enabling fine-grained diagnosis of where model failures originate.

3.2 Data Curation

We construct FBI-Bench through a 3-stage pipeline: **Image sourcing**, **QA Annotation**, and **Filterings**. Pipeline is designed to ensure each retained sample is visually grounded, human-easy, and diagnostic of shared VLM failures. Detailed annotation protocol is provided in Appendix E

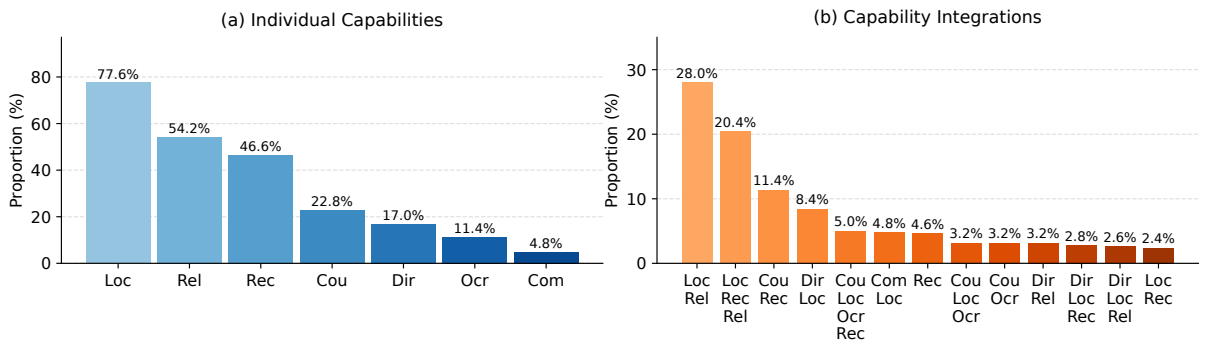


Figure 4: **Capability distribution in FBI-Bench.** (a) Proportion of individual capabilities. The total exceeds 100% because samples may contain multiple capabilities. (b) Proportion of capability integrations (unique capability combinations per sample), which sums to 100%. Abbreviations: Dir (Direction), Loc (Location), Rec (Recognition), Rel (Relation), Cou (Counting), Com (Comparison), and Ocr (OCR).

Image sourcing. In the first stage, 10 annotators collect candidate images from public internet sources and prior benchmarks (COCO-style natural-image datasets and VCR-style visual reasoning datasets). Annotators filter out images with poor visual quality, unclear source information, privacy concerns, redistribution restrictions, or content requiring specialized domain knowledge.

QA Annotation. In the second stage, annotators write or refine free-form QA pairs for the retained images. Each question must be answerable from the image alone. Annotators provide a answer and capability labels corresponding to the minimal skills required to solve the sample. Candidates are kept only if they are human-easy, unambiguous, and grounded in direct visual evidence.

Filterings. In the third stage, we remove duplicate, ambiguous, irrelevant, incorrectly labeled, or non-diagnostic samples through multi-round checking. To further identify systematic weaknesses across current VLMs, we apply model-based filtering using multiple evaluated models. Samples solved by at least one evaluated model are excluded, while only those failed by all evaluated models are retained. This process reduces model-specific noise and helps FBI-Bench isolate shared limitations in fundamental visual perception.

Data statistics. After filtering, FBI-Bench contains 500 samples spanning seven foundational skills and 13 capability integrations. This enables analysis of both isolated perceptual failures and failures that emerge when multiple skills must be coordinated. The empirical distribution is shown in **Fig. 4(b)**. To verify that the benchmark is human-solvable, we evaluate participants from three education levels: elementary school, middle school, and high school, achieving 81.2%, 87.8%, and 96.4% accuracy, respectively. These results indicate that FBI-Bench is well-posed for humans while remaining highly challenging for current VLMs.

4 Experiments

4.1 Experimental Setup

We evaluate 14 state-of-the-art vision–language models (VLMs), covering both proprietary frontier systems and large open-weight models, as shown in **Tab. 1**. All evaluations are conducted in a strict zero-shot setting, without task-specific fine-tuning, external tools, retrieval augmentation, or chain-of-thought prompting. During inference, we use greedy decoding with temperature $= 0.0$ and require models to provide direct answers only.

We evaluate models under an **open-ended generation** setting rather than the traditional multiple-choice question (MCQ) format. This setting reduces the chance that models exploit option biases, answer priors, or partial linguistic cues (Yu et al., 2023; Li et al., 2023; Yue et al., 2024). For each sample, the model receives the image and the corresponding free-form question. Model responses are scored using an LLM-as-a-judge protocol selected through meta-evaluation against human annotations. The judge evaluates semantic consistency between the model prediction and the reference answer, assigns a score from 0.0 to 1.0, and treats scores ≥ 0.5 as correct. Further details on prompts, decoding hyperparameters, judge selection, and validation are provided in **Appendix C**.

Table 1: Zero-shot performance of 14 MLLMs on FBI-Bench across core abilities (Accuracy %). *These models are used for model-based filtering during dataset construction; therefore, resulting in 0% accuracy across abilities.*

Model	Accuracy (%)							
	Location	Relation	Recognition	Counting	Direction	OCR	Comparison	Overall
Human Performance								
Primary School Student	89.3	86.7	81.5	77.4	73.2	69.1	61.8	81.2
Middle School Student	94.1	92.6	87.4	84.8	82.9	79.6	76.3	87.8
High School Student	98.2	97.6	96.8	95.9	94.7	93.8	91.5	96.4
Proprietary MLLMs								
GPT-4o(OpenAI et al., 2024a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GPT-5-Mini(Singh et al., 2025)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude-Haiku-4.5(Anthropic, 2025b)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude-Sonnet-4.5(Anthropic, 2025a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gemini-2.5-Flash(Gemini Team, 2025b)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gemini-2.0-Flash(Google, 2025)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Open-weight MLLMs								
360VL-70B(Qihoo 360 and 360CVGroup, 2024)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
InternVL2.5-78B(Chen et al., 2025b)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
InternVL3-78B(Zhu et al., 2025)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kimi-VL-A3B-Inst.(Team et al., 2025)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-VL-72B-Inst.(Bai et al., 2025b)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-8B-Inst.(Bai et al., 2025a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Molmo-72B-0924(Deitke et al., 2024)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-OneVision-7B(Li et al., 2024a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Benchmarks	Failure Roots Frequency				
	Low-Level Visual Confounders	Spatial Imagination	Prior-Knowledge Overwriting	Soft Cognitive Variance	Others
FBI-Lite	- Small Objects: 4.6% - Occlusion: 1.0% - High Density: 7.2%	- Viewpoint Shift: 11.0% - Pose/Orientation: 8.6% - Spatial Relation: 10.0%	- Superficial Grounding: 2.2% - Evidential Perception: 2.0%	- Counting Variance: 12.2% - Attribute Confusion: 14.4% - Text Misreading: 10.0%	16.8%
FBI	- Small Objects: 3.6% - Occlusion: 1.4% - High Density: 7.8%	- Viewpoint Shift: 24.0% - Pose/Orientation: 10.6% - Spatial Relation: 22.4%	- Superficial Grounding: 22.2% - Evidential Perception: 6.4%	- Counting Variance: 0.0% - Attribute Confusion: 0.0% - Text Misreading: 0.0%	1.6%

Table 2: Contrastive profile of FBI-Lite and FBI to identify the roots of the SOTA performance collapse.

4.2 Main Results

Tab. 1 reports the performance of 14 state-of-the-art MLLMs on the seven foundational capabilities in FBI-Bench. Despite strong results on existing multimodal benchmarks, all evaluated models obtain 0.0% overall accuracy under our standardized free-form evaluation protocol (Sec. 4.1), failing consistently across recognition, localization, relation reasoning, counting, direction understanding, OCR, and comparison.

To ensure reliability, we manually inspect representative predictions and verify that the reported results accurately reflect model behavior under the defined protocol. In contrast, human participants achieve high performance under the same evaluation setting, with primary, middle, and high school students reaching 81.2%, 87.8%, and 96.4% overall accuracy, respectively.

These findings reveal a substantial gap between benchmark-level performance and genuine fundamental visual understanding, suggesting that current MLLMs still lack reliable mastery of basic perceptual and compositional skills.

4.3 Failure Analysis

To test whether the 0.0% floor reflects genuine perceptual limitations rather than adversarial image selection, we construct **FBI-Lite**, a 500-sample subset from the pre-filtered pool excluded from FBI because at least one model solved them. This provides a controlled baseline for direct comparison under the same failure taxonomy. Tab. 2 compares the two sets across five categories: low-level visual confounders, spatial imagination, prior-knowledge overwriting, soft cognitive variance, and others. Detailed definitions and qualitative analysis are pro-

Table 3: Ablation 1: Effect of explicit spatial interventions across 13 complex capability integrations. Accuracy is reported in %.

Model	Modality	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Base (No Intervention)															
LLaVA-OneVision-7B(Li et al., 2024a)	None	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-8B-Inst.(Bai et al., 2025a)	None	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dense Grounding															
LLaVA-OneVision-7B(Li et al., 2024a)	BBox in text	1.4	2.9	0.0	2.4	0.0	4.2	4.3	0.0	0.0	0.0	7.1	7.7	8.3	2.2
	BBox in image	1.4	2.9	0.0	0.0	8.0	8.3	0.0	0.0	0.0	0.0	7.1	0.0	8.3	2.2
Qwen3-VL-8B-Inst.(Bai et al., 2025a)	BBox in text	0.0	2.9	0.0	4.8	4.0	0.0	8.7	0.0	0.0	0.0	0.0	0.0	0.0	1.6
	BBox in image	0.0	1.0	1.8	0.0	12.0	0.0	4.3	6.3	0.0	0.0	0.0	7.7	0.0	1.6
Oracle Attention															
LLaVA-OneVision-7B(Li et al., 2024a)	BBox in text	1.4	2.9	1.8	2.4	0.0	4.2	4.3	0.0	0.0	0.0	7.1	0.0	8.3	2.2
	BBox in image	2.1	2.9	1.8	0.0	8.0	0.0	0.0	0.0	0.0	6.3	0.0	0.0	8.3	2.2
Qwen3-VL-8B-Inst.(Bai et al., 2025a)	BBox in text	1.4	0.0	5.3	2.4	0.0	4.2	4.3	0.0	0.0	0.0	0.0	0.0	8.3	1.8
	BBox in image	1.4	1.0	3.5	2.4	4.0	4.2	8.7	6.3	0.0	0.0	7.1	0.0	0.0	2.4

Table 4: Ablation 2: Evaluation results on MLLMs with or not latent thinking regarding each capability integration. Accuracy is presented in %.

Model	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Gemini-2.5-Flash(Gemini Team, 2025b)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gemini-2.5-Pro(Gemini Team, 2025a)	4.3	5.9	3.5	4.8	7.1	4.2	0.0	18.8	37.5	0.0	4.0	0.0	16.7	6.0
Qwen3-VL-8B-Inst.(Bai et al., 2025a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-8B-Thinking(Bai et al., 2025a)	5.0	2.0	1.8	4.8	12.0	8.3	8.7	21.4	43.8	6.3	0.0	7.7	8.3	6.0
Kimi-VL-A3B-Inst.(Team et al., 2025)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Kimi-VL-A3B-Thinking(Team et al., 2025)	1.7	8.8	7.0	0.0	8.0	8.3	13.0	18.8	10.7	0.0	0.0	30.8	0.0	8.0

vided in **Appendix F** and **Appendix G**. Three findings emerge. First, low-level image difficulty is not the main factor: visual confounders appear at similar rates in both sets, suggesting FBI is not simply composed of harder images. Second, spatial imagination is the dominant failure source: viewpoint-shift errors increase from 11.0% to 24.0%, while spatial-relation errors rise from 10.0% to 22.4%, showing that universally failed samples require integrating spatial evidence across reference frames. Third, FBI-Lite successes often rely on shortcut reasoning rather than genuine understanding. Superficial grounding increases from 2.2% to 22.2% in FBI, whereas soft cognitive variance drops to 0.0%, indicating that removing shortcut-solvable samples exposes the core perceptual limitation. The full curation funnel is reported in **Appendix B.2.2**.

4.4 Ablation Studies

Does Direct Visual Intervention Resolve Spatial Blindness? We test if spatial guidance can mitigate spatial failures by adding bounding boxes from Grounding DINO (Liu et al., 2024b). We use either *Dense Grounding*, which marks all detected objects, or *Oracle Attention*, which highlights only query-relevant objects. The guidance is provided as either text coordinates or visual overlays. **Tab. 3** shows that these yield limited improvement. Even oracle boxes fail to substantially improve spatial-relation tasks such as {Loc, Rel} and {Dir, Loc, Rel}, whose accuracies remain below 10%. Together with the Grad-CAM evidence in **Appendix D.11**, suggesting that models can access object locations but fail to compose them into spatial relations. The bottleneck is not merely localization, but the lack of reliable spatial grounding.

> ***Finding 1.*** VLMs can often locate coordinates while remaining unable to synthesize them into a coherent logical structure.

Does increasing the latent thinking budget re-

Table 5: Ablation 3: Evaluation results on various vision encoders and LLMs regarding each capability integration. Accuracy is presented in %.

Model	Image Encoder	LLM	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Llava-1.5-7B(Liu et al., 2023a)	CLIP ViT-L/14(Radford et al., 2021)	Vicuna-7B(Zheng et al., 2023)	3.6	7.8	1.8	11.9	0.0	4.2	13.0	0.0	0.0	37.5	0.0	0.0	0.0	5.8
Llava-v1.6-7B(Liu et al., 2024a)	CLIP ViT-L/14(Radford et al., 2021)	Vicuna-7B(Zheng et al., 2023)	4.3	7.8	0.0	0.0	0.0	20.8	4.4	0.0	0.0	18.8	7.1	0.0	8.3	5.0
Llava-v1.6-7B(Liu et al., 2024a)	CLIP ViT-L/14(Radford et al., 2021)	Mistral-7B(Jiang et al., 2023)	5.7	6.9	3.5	2.4	0.0	4.2	8.7	0.0	0.0	18.8	0.0	0.0	8.3	5.0
Idefics2-8B(Laurençon et al., 2024, 2023)	SigLIP(Zhai et al., 2023; Tschannen et al., 2025)	Mistral-7B(Jiang et al., 2023)	15.0	17.6	0.0	4.8	0.0	12.5	4.4	0.0	0.0	31.2	7.1	7.7	8.3	10.6
InternVL2-8B(Chen et al., 2024c)	InternViT-6B(Chen et al., 2024c)	InternLM2-7B(Chen et al., 2024c)	4.3	2.9	0.0	0.0	4.0	16.7	4.4	12.5	0.0	6.2	0.0	0.0	8.3	3.8

Table 6: Ablation 3: Evaluation results on SigLIP-based MLLMs (Zhai et al., 2023; Tschannen et al., 2025) regarding each capability integration. Accuracy is presented in %.

Model	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Idefics2-8B(Laurençon et al., 2024, 2023)	15.0	17.6	0.0	4.8	0.0	12.5	4.3	0.0	0.0	31.2	7.1	7.7	8.3	10.6
Gemma-3-1B(Team, 2025)	12.1	16.7	10.5	0.0	8.0	20.8	0.0	0.0	0.0	6.2	0.0	7.7	8.3	10.0
Gemma-3-4B(Team, 2025)	13.6	21.6	1.8	0.0	8.0	12.5	13.0	6.2	0.0	37.5	0.0	7.7	8.3	11.8
Gemma-3-12B(Team, 2025)	20.7	18.6	3.5	2.4	0.0	12.5	17.4	0.0	0.0	25.0	7.1	15.4	0.0	13.0
Gemma-3-27B(Team, 2025)	18.6	24.5	8.8	2.4	0.0	16.7	4.3	6.2	0.0	18.8	0.0	7.7	0.0	13.4
Paligemma-3B(Beyer* et al., 2024)	32.9	39.2	1.8	16.7	4.0	41.7	8.7	0.0	0.0	43.8	14.3	30.8	0.0	24.0

solve visual perception errors? **Tab. 4** shows that increasing the latent thinking budget does not reliably improve fundamental visual perception, especially on spatial and geometric tasks. When the vision encoder misses key structural evidence, additional reasoning cannot recover the correct answer; instead, models often produce coherent explanations based on incorrect visual premises. This supports a clear perception–reasoning gap and is consistent with prior observations that more computation may not help when the bottleneck is perception (Tian et al., 2025). We observe limited gains mainly on OCR-heavy tasks, such as {Cou, OCR}, where Qwen3-VL reaches 43.8%. This suggests that current VLMs benefit more when visual information can be directly converted into symbolic text, but remain weak when success requires precise spatial grounding or geometric understanding.

> ***Finding 2.*** Increasing the latent thinking budget often turns misperceived visual inputs into plausible but incorrect reasoning chains.

How much does the vision encoder affect spatial grounding? **Tab. 5** shows that vision encoder design strongly affects spatial grounding. SigLIP-based Idefics2-8B (Laurençon et al., 2024, 2023; Zhai et al., 2023) achieves 15.0% on {Loc, Rel} and 17.6% on {Loc, Rec, Rel}, substantially outperforming CLIP-based models such as LLaVA-v1.6-Mistral (Radford et al., 2021; Liu et al., 2024a), suggesting that SigLIP preserves fine-grained local structure more effectively than encoders optimized mainly for global image–text alignment. The trend becomes clearer in **Tab. 6**, where SigLIP-based Gemma-3 (Team, 2025) and PaliGemma (Beyer* et al., 2024) achieve the strongest results. In particular, PaliGemma-3B reaches 24.0% overall accuracy, outperforming much larger 72B–78B models, while Gemma-3-4B achieves 21.6% on the challenging {Loc, Rec, Rel} setting. However, even these stronger models remain weak on OCR-counting compositions such as {Cou, OCR}, indicating that current architectures still struggle to jointly integrate textual recognition, counting, and spatial grounding within a single reasoning step.

> ***Finding 3.*** Vision encoder quality and clean vision–language alignment matter more than sheer model scale for fundamental multimodal capability.

5 Conclusion

FBI-Bench evaluates the fundamental visual capabilities of vision-language models. Despite strong performance on complex reasoning benchmarks, current VLMs still fail on elementary tasks such as spatial orientation, counting, and local attribute recognition. We further show that common improvement strategies provide limited gains, while stronger vision encoders and better vision–language alignment are substantially more effective. These results highlight visual perception as a major open challenge for multimodal intelligence.

6 Limitations

FBI-Bench is designed as a diagnostic benchmark for fundamental visual skills rather than a comprehensive evaluation of all multimodal abilities. Its failure-focused construction intentionally emphasizes examples that expose persistent weaknesses in current VLMs, providing a targeted stress test for visual grounding and perception. While future advances in architectures and vision encoders may reduce the observed gap, we hope **FBI-Bench** serves as a useful benchmark for tracking progress in fundamental multimodal understanding.

References

Samira Abnar, Harshay Shah, Dan Busbridge, Alaaeldin Mohamed Elnouby Ali, Josh Susskind, and Vimal Thilak. 2025. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models.

Sammy Ahmed, Dominic Kelly, Nicholas Waters, and Natasha Chaku. 2024. Executive functioning.

Abigail A. Allen, R. Alex Smith, Matthew K. Burns, and Erica S. Lembke. 2024. Early academic and behavior skills as predictors of later mathematics achievement. *Psychology in the Schools*, 61(7):3010–3025.

Anthropic. 2025a. Claude 4.5 sonnet system card.

Anthropic. 2025b. Claude haiku 4.5 system card.

Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. 2025. Benchmarking and mitigating mcqa selection bias of large vision-language models.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025a. Qwen3-vl technical report.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025b. Qwen2.5-vl technical report.

Horace Barlow. 1990. Conditions for versatile learning, helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30(11):1561–1571. Optics Physiology and Vision.

Lucas Beyer*, Andreas Steiner*, André Susano Pinto*, Alexander Kolesnikov*, Xiao Wang*, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai*. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. 2025a. Bring reason to vision: Understanding perception and reasoning through model merging.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Jiwan Chung, Neel Joshi, Pratyusha Sharma, Youngjae Yu, and Vibhav Vineet. 2026. What mllms learn about when they learn about multimodal reasoning: Perception, reasoning, or their integration?

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models.

Adele Diamond. 2020. Executive functions. *Handbook of clinical neurology*, 173:225–240.

Naomi P. Friedman and Akira Miyake. 2017. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, 86:186–204. Is a "single" brain model sufficient?

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025a. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. 2025b. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. 2025c. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive.

Gemini Team. 2025a. Gemini 2.5 flash system card.

Gemini Team. 2025b. Gemini 2.5 flash system card.

Simon Ging, María A. Bravo, and Thomas Brox. 2024. Open-ended vqa benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy.

Google. 2025. Gemini 2.0 flash model card.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Samuel Greiff, Sascha Wüstenberg, Thomas Goetz, Mari-Pauliina Vainikainen, Jarkko Hautamäki, and Marc Bornstein. 2015. A longitudinal study of higher-order thinking skills: working memory and fluid reasoning in childhood enhance complex problem solving in adolescence. *Frontiers in Psychology*, 6:1060.

Xiaotian Han, Yiqi Wang, Bohan Zhai, Quanzeng You, and Hongxia Yang. 2024. Coco is "all" you need for visual instruction fine-tuning.

Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, Jie Fu, Ziwei Liu, Jinwoo Shin, Kimin Lee, Mantas Mazeika, Long Phan, George Ingebretsen, Adam Khoja, Cihang Xie, Olawale Salaudeen, Matthias Hein, Kevin Zhao, Alexander Pan, David Duvenaud, Bo Li, Steve Omohundro, Gabriel Alfour, Max Tegmark, Kevin McGrew, Gary Marcus, Jaan Tallinn, Eric Schmidt, and Yoshua Bengio. 2025. A definition of agi.

Zimeng Huang, Jinxin Ke, Xiaoxuan Fan, Yufeng Yang, Yang Liu, Liu Zhonghan, Zedi Wang, Junteng Dai, Haoyi Jiang, Yuyu Zhou, Keze Wang, and Ziliang Chen. 2025. Mm-opera: Benchmarking open-ended association reasoning for large vision-language models.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Botian Jiang, Lei Li, Xiaonan Li, Zhaowei Li, Xiachong Feng, Lingpeng Kong, Qi Liu, and Xipeng Qiu. 2024. Understanding the role of llms in multimodal evaluation benchmarks.

Aditya Kanade and Tanuja Ganu. 2025. Do you see me : A multidimensional benchmark for evaluating visual perception in multimodal llms.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models?

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023. Seed-bench-2: Benchmarking multimodal large language models.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024c. A survey on benchmarks of multimodal large language models.

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. 2024d. A survey on multimodal benchmarks: In the era of large ai models.

Shengzhi Li, Rongyu Lin, and Shichao Pei. 2024e. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 14188–14200. Association for Computational Linguistics.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. 2025a. On the perception bottleneck of vlms for chart understanding.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Yesheng Liu, Hao Li, Haiyu Xu, Baoqi Pei, Jiahao Wang, Mingxuan Zhao, Jingshu Zheng, Zheqi He, JG Yao, Bowen Qin, Xi Yang, and Jiajun Zhang. 2025b. Beyond multiple choice: Verifiable openqa for robust vision-language rft.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player?

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024d. Mmbench: Is your multi-modal model an all-around player?

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. 2025a. Towards robust mathematical reasoning.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. 2025b. Towards robust mathematical reasoning.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants.

Akira Miyake and Naomi Friedman. 2012. The nature and organization of individual differences in executive functions: Four general conclusions. *Current directions in psychological science*, 21:8–14.

Alhassan Mumuni and Fuseini Mumuni. 2025. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches.

Camila Peres Nogues and Beatriz Vargas Dorneles. 2021. Systematic review on the precursors of initial mathematical performance. *International Journal of Educational Research Open*, 2:100035.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers,

Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. Gpt-4o system card.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024b. Gpt-4 technical report.

Qihoo 360 and 360CVGroup. 2024. 360vl-70b: Open source large multi-modal model.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Ju-

lian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, Vatsal Raina, Hanyi Xiong, Vishaal Udandarao, Jingyi Lu, Shiyang Chen, Sam Purkis, Tianshuo Yan, Wenye Lin, Gyungin Shin, Qiaochu Yang, Anh Totti Nguyen, David I. Atkinson, Aaditya Baranwal, Alexandru Coca, Mikah Dang, Sebastian Dziadzio, Jakob D. Kunz, Kaiqu Liang, Alexander Lo, Brian Pulfer, Steven Walton, Charig Yang, Kai Han, and Samuel Albanie. 2025. Zerobench: An impossible visual benchmark for contemporary large multimodal models.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie,

Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2025. Openai gpt-5 system card.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yangfan He, Kuan Lu, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2025. Mtvqa: Benchmarking multilingual text-centric visual question answering.

Gemma Team. 2025. Gemma 3.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. 2025. Kimi-vl technical report.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. 2025. More thought, less accuracy? on the dual nature of reasoning in vision-language models.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning. https://github.com/huggingface/trl.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Llms may perform mcqa by selecting the least incorrect option.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency.

Xinhe Wang, Jin Huang, Xingjian Zhang, Tianhao Wang, and Jiaqi W. Ma. 2026. Your reasoning benchmark may not test reasoning: Revealing perception bottleneck in abstract reasoning benchmarks.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-bench: A benchmark for general-purpose foundation models on low-level vision.

X.AI. 2024. Grok-1.5 vision preview.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization.

Yan Yang, Dongxu Li, Haoning Wu, Bei Chen, Liu Liu, Liyuan Pan, and Junnan Li. 2025. ProBench: Judging multimodal foundation models on open-ended multi-domain expert tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10883–10892, Vienna, Austria. Association for Computational Linguistics.

Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K. Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, Wenhao Wu, and Dacheng Tao. 2025. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi.

Jie Ying, Zihong Chen, Zhefan Wang, Wanli Jiang, Chenyang Wang, Zhonghang Yuan, Haoyang Su, Huanjun Kong, Fan Yang, and Nanqing Dong. 2025. Seedbench: A multi-task benchmark for evaluating large language models in seed science.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and Xiangyu Yue. 2025. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark.

Giselle Zeno, Nour Jedidi, and Steven R. Gomez. 2025. Choosing 'right' from wrong: A closer look at selection bias in spatial multiple-choice questions in large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 535–544.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training.

Hao Zhang, Chen Li, and Basura Fernando. 2025a. Mitigating easy option bias in multiple-choice question answering.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2025b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. 2024b. Q-bench+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, Fei Luo, Xiaohua Chen, Xiaoshuai Hao, Hehan Li, Andi Zhang, Wenxuan Wang, Kaiyan Zhang, Guoli Jia, Lingling Li, Zhiwu Lu, Yang Lu, and Yike Guo. 2025a. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models.

Zhihan Zhou, Feng Hong, Jiaan Luo, Jiangchao Yao, Dongsheng Li, Bo Han, Ya Zhang, and Yanfeng Wang. 2025b. Learning to instruct for visual instruction tuning.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He,

Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models.

FBI-Bench: Fundamental but Impossible - Are We Overrating VLMs' Capabilities?

Supplementary Material

Table of Contents in Appendix

A Why FBI-Bench matters 19
A.1 Why "Fundamental"? 19
A.2 Why "Impossible"? 20
A.3 Why Free-Form Evaluation? 20
A.4 Why A Lightweight Benchmark? 21

B Additional Discussion 22
B.1 Comparison with Existing Multimodal Benchmarks 22
B.2 More About FBI-Bench 24
B.2.1 Ethical considerations 24
B.2.2 A benchmark is most useful when it becomes obsolete 25
B.2.3 Release and community use 25
B.3 Future Work 26

C Experimental Setting 27
C.1 Prompts 27
C.2 Evaluation Protocol and Meta-Evaluation 27
C.3 Hyperparameters 28
C.4 Models 28
C.5 Human Baseline 28

D Additional Ablation Studies 31
D.1 Ablation 1: The Zero-Consistency Dilemma and Visual Lottery 31
D.2 Ablation 2: Open-ended Evaluation vs. The Multiple-Choice Illusion 31
D.3 Ablation 3: Can Prompt Rethinking Trigger Visual Self-Correction? (Supplementary) 31
D.4 Ablation 4: The Alignment Tax on Visual Grounding (Supplementary) 31
D.5 Ablation 5: Inverse Scaling and Semantic Overshadowing (Supplementary) . 33
D.6 Ablation 6: Decoding Settings and Representation Collapse (Supplementary) . . 33
D.7 Ablation 7: Resolution Scaling and Selective Spatial Gain 33
D.8 Ablation 8: Does data pretraining affect the fundamental skills of VLMs? 35
D.9 Ablation 9: Post-training alignment and fundamental visual skills 35
D.10 Ablation 10: Scaling the language backbone in VLMs 35

E Data Annotation and Curation Protocol 37
E.1 Overview 37
E.2 Annotator Recruitment and Responsibilities 37
E.3 Stage 1: Image Collection and Visual Filtering 38
E.4 Stage 2: Question–Answer Annotation and Capability Labeling 39
E.5 Stage 3: Multi-Round Filtering and Final Selection 41
E.6 Ambiguity Resolution 42
E.7 Copyright, Privacy, and Source-Use Policy 42
E.8 Annotation Format and JSON Schema 43
E.9 Human Validation 43
E.10 Release Format and Community Use 43

F Qualitative Analysis 48

A Why FBI-Bench matters

A.1 Why "Fundamental"?

From Perception to Cognition: In cognitive psychology, higher-order abilities are typically understood as emerging from more basic control functions rather than in isolation: executive functions (Diamond, 2020; Ahmed et al., 2024) such as working memory, inhibitory control (Miyake and Friedman, 2012), and cognitive flexibility support reasoning, planning, and complex problem solving (Greiff et al., 2015; Friedman and Miyake, 2017). Longitudinal and developmental studies further show that early foundational capacities predict later higher-order performance, including fluid reasoning, academic achievement, and complex problem solving (Greiff et al., 2015; Nogues and Dorneles, 2021; Allen et al., 2024). In other words, higher-order competence is only as reliable as the lower-level skills that support it. A similar decomposition is increasingly visible in multimodal model. Broad VLM evaluation benchmarks (Fu et al., 2025a; Liu et al., 2024c; Yue et al., 2025; Yu et al., 2023) explicitly separate perception from cognition, reflecting the view that strong multimodal behavior depends on both layers rather than on reasoning alone (Yue et al., 2024; Chen et al., 2024b). Across multiple settings, these works (Liu et al., 2025a; Chen et al., 2025a) show that perception often acts as a bottleneck for downstream reasoning, that perception and reasoning are functionally distinct but interdependent, and that many apparent reasoning failures are in fact caused by weak perception or poor perception–reasoning integration rather than reasoning alone (Yuan et al., 2025; Chung et al., 2026; Wang et al., 2026). Therefore, from both the human-learning perspective and the multimodal-learning perspective, the same principle holds: before interpreting success on complex tasks as evidence of robust intelligence, we must first verify that the model has mastered the fundamental skills.

Failures on easy cases matter because they are often the failures users notice first. In real-world interaction, trust is often undermined not by failure on an exceptionally difficult problem, but by failure on a question that appears trivial to humans. This is exactly why checking fundamental skills is necessary. Recent work shows that strong performance on challenging multimodal benchmarks does not necessarily imply reliable basic competence: models may succeed on complex tasks while still failing human-easy perceptual ones (Fu et al., 2024; Chen et al., 2024b). More broadly, benchmark design itself can mask such weaknesses when models benefit from shortcuts or when perception is not evaluated separately from higher-level cognition (Roberts et al., 2025; Wu et al., 2024; Zhang et al., 2024b). Therefore, evaluating fundamental skills is essential for aligning model assessment with the kinds of simple but visible failures that most directly shape user trust.

A credible path toward AGI requires reliability on easy tasks, not only success on hard ones. Recent excitement around advanced AI systems (Gemini Team, 2025a; Anthropic, 2025b; Singh et al., 2025; Bai et al., 2025a) often centers on their performance on highly difficult benchmarks (Yue et al., 2025, 2024), leading to speculation that general intelligence may be within reach. However, strong results on difficult tasks are not sufficient evidence of trustworthy general intelligence. A genuinely reliable AGI should not merely solve spectacular problems at the high end of the difficulty spectrum (Luong et al., 2025a,b); it should also behave consistently on elementary cases that lie well within its competence (Yuan et al., 2025; Chung et al., 2026). If a system can answer a difficult puzzle yet fails a basic visual question, then its intelligence is impressive but not dependable. In this sense, checking fundamental skills serves as a reliability test: before we interpret high-end performance as evidence of general competence, we must first verify that the system is stable on the basics.

As in human development, mastery of complex tasks should be grounded in mastery of basic ones first. Evaluating fundamental skills first is necessary because competence in complex tasks typically grows out of competence in simpler, prerequisite ones. In human learning (Miyake and Friedman, 2012; Nogues and Dorneles, 2021),unknown11, foundational abilities are acquired (Greiff et al., 2015; Diamond, 2020) before higher-order performance, and early skills serve as the basis for later competence and proficiency (Friedman and Miyake, 2017; Allen et al., 2024). For that reason, education research commonly uses diagnostic assessment to identify whether prerequisite skills are already in place before interpreting failure on more advanced tasks. The same principle should apply to multimodal models: before claiming that a model fails at "reasoning," we should first verify whether it can reliably perform the underly-

ing perceptual and elementary compositional operations that the task depends on. This is important not only for fair evaluation but also for diagnosis, because poor performance on a complex benchmark may reflect a failure in perception rather than a true breakdown in reasoning. Therefore, checking fundamental skills first provides a more principled way to locate the real source of error and to guide the development of stronger multimodal systems.

A.2 Why "Impossible"?

A reliable multimodal system should not only succeed on prestigious, high-complexity reasoning tasks; it should also remain dependable on elementary perceptual judgments (in **Sec. A.1**). Recent evaluation work likewise suggests that perception is not a peripheral skill, but a prerequisite for higher-level multimodal reasoning and reliable visual understanding (Chen et al., 2024b; Zhou et al., 2025a; Li et al., 2024c,d). From this perspective, **impossible** is not a provocative label, but a deliberate way to target the regime in which current models still fail on the very foundations that advanced reasoning presupposes.

Impossibility is necessary to reveal systematic limitations rather than model-specific errors. If only one model fails on an example while other strong models solve it, the failure can plausibly be attributed to idiosyncratic factors such as architecture choice, training data mixture, alignment strategy, or decoding behavior (Li et al., 2025; Zhang et al., 2024a). Such cases are useful for model comparison, but they do not isolate a field-level bottleneck (Li et al., 2024c,d). By contrast, when diverse frontier and open-weight models all fail on the same human-easy sample, the failure is much more likely to reflect a shared missing capability in current VLMs. In this way, the benchmark is designed to surface limitations that are stable across architectures and training pipelines, rather than incidental mistakes of a single system.

Impossibility is necessary to preserve benchmark headroom and resist rapid saturation. Many multimodal benchmarks (Liu et al., 2024c; Fu et al., 2025a; Li et al., 2023; Ying et al., 2025) quickly saturate as models improve, after which high scores provide little resolution for measuring meaningful progress (Yue et al., 2025; Roberts et al., 2025). A common response has been to move toward increasingly difficult, composite, or knowledge-intensive tasks (Burgess et al., 2025; Lu et al., 2024; Luong et al., 2025b). However, this shift introduces a different problem: strong performance on hard benchmarks does not guarantee reliable mastery of basic skills (Fu et al., 2024; Wu et al., 2024). In fact, greater task complexity can obscure failures in elementary perception, making it harder to determine whether a model genuinely understands the visual input or merely succeeds through shortcut strategies, answer selection and recognition, or partial cues (Yu et al., 2023; Li et al., 2023; Yue et al., 2024; Chen et al., 2024b; Wang et al., 2024). An impossible-yet-fundamental benchmark avoids this trade-off. Because the tasks remain simple and human-easy, any measurable gain is more interpretable: it is more likely to reflect real improvement in perceptual grounding, rather than benchmark artifacts or superficial reasoning heuristics.

In short, **impossible is a methodological requirement, not a rhetorical flourish**. It ensures that FBI-Bench functions as a diagnostic benchmark for foundational weaknesses shared across current VLMs, rather than as a leaderboard that mostly captures model-specific variance. At the same time, it prevents the benchmark from becoming uninformative shortly after release. A benchmark that is impossible for current models—while still unambiguous and easy for humans—is precisely the kind of test needed to measure whether future systems have truly acquired the fundamental perceptual skills on which reliable multimodal intelligence depends.

A.3 Why Free-Form Evaluation?

Prior evidence: A growing body of recent work (Wang et al., 2024; Yao et al., 2025; Kanade and Ganu, 2025; Jiang et al., 2024) suggests that multiple-choice evaluation can substantially overestimate a model's actual capability. One line of work (Zeno et al., 2025; Yao et al., 2025) shows that, under MCQ settings, models may succeed by exploiting *selection bias* and answer-option artifacts, rather than by truly solving the underlying problem. In particular, prior studies (Yu et al., 2023; Chen et al., 2024b) show that models can benefit from option-order bias, answer-position preferences, easy-option shortcuts (Atabuzzaman et al., 2025; Zhang et al., 2025a), or even select the *least incorrect* option instead of a genuinely grounded one (Wang et al., 2024). Another line of work shows that strong performance on existing multi-

modal benchmarks does not necessarily imply true multimodal understanding (Yue et al., 2025; Jiang et al., 2024; Yang et al., 2025). In response to these concerns, several recent efforts have moved toward open-ended evaluation (Liu et al., 2024c; Yue et al., 2024; Huang et al., 2025), arguing that free-form generation reduces guessability and provides a more faithful measure of genuine multimodal reasoning and understanding (Ging et al., 2024; Yao et al., 2025; Liu et al., 2025b). Taken together, these prior findings suggest that while MCQ is convenient, it is not an ideal format when the goal is to assess whether a VLM truly understands the visual input.

Our evidence: Our observations on FBI-Bench support the same conclusion. First, as shown in **Fig. 2(b)**, a model may choose the correct answer when it appears among options, yet fail on the same example under strict free-form evaluation. This suggests that some MCQ success comes from answer recognition rather than genuine understanding. To examine this further, we compare MCQ and free-form evaluation on 100 randomly sampled FBI-Bench instances across five models (**Tab. 13**). All models improve under MCQ. Notably, models scoring 0.0% in free-form, such as Qwen3-VL-8B-Instruct and Kimi-VL-A3B-Instruct, still reach 21% and 30% MCQ accuracy, indicating gains more consistent with option recognition or guessing than real visual understanding. The MCQ advantage also increases with scale in the InternVL3.5 family: the 78B model reaches 55% under MCQ, more than doubling the 8B model's 23%. This pattern suggests that scaling may improve the ability to exploit answer options more than it improves visual grounding, aligning with prior findings that larger language backbones can amplify linguistic priors without strengthening visual understanding (Ging et al., 2024; Yao et al., 2025; Liu et al., 2025b).

Taken together, prior work and our experiments lead to the same conclusion: MCQ can inflate performance and obscure whether success reflects genuine perception or option selection. Our goal is not simply to make the benchmark harder, but more diagnostic. If the central question is whether current VLMs possess reliable foundational visual skills, the evaluation must reduce shortcut opportunities and depend on actual perception. Free-form evaluation better satisfies this requirement and thus provides a more faithful estimate of current VLM capabilities. Since FBI-Bench aims to diagnose *fundamental* visual skills, we adopt *free-form evaluation* as the primary protocol.

A.4 Why A Lightweight Benchmark?

Modern multimodal benchmarks (Yue et al., 2024; Fu et al., 2025c; Zhang et al., 2025b) are increasingly large-scale, often containing tens of thousands or even millions of samples. While such scale is useful for broad coverage, it also makes evaluation increasingly expensive, especially when testing many frontier and open-source models under free-form generation (Yue et al., 2025; Ying et al., 2025; Li et al., 2023). This creates a practical limitation: large benchmarks require substantial computation and cost, which makes repeated diagnosis, controlled comparison, and rapid iteration harder than they need to be.

More importantly, larger scale does not necessarily translate into deeper insight. In many cases, massive benchmarks mainly provide aggregate scores, while offering limited visibility into the systematic errors that models continue to share (Yu et al., 2023; Chen et al., 2024b; Fu et al., 2025a). For our purpose, this is a critical limitation. **FBI-Bench** is not designed to maximize size, but to maximize diagnostic value: each retained sample is human-easy, targets fundamental skills, and is selected because it reveals a failure that persists across diverse models rather than a model-specific mistake.

For this reason, we revisit benchmark design from a different perspective. Instead of building another large-scale benchmark, we keep FBI-Bench compact and lightweight, with 500 carefully curated samples. This makes evaluation far more practical, while still exposing a wide range of systematic limitations in current frontier and open-source VLMs. In this sense, the value of FBI-Bench lies not in raw scale, but in its ability to reveal important and actionable weaknesses with much higher efficiency.

B Additional Discussion

B.1 Comparison with Existing Multimodal Benchmarks

FBI-Bench is designed to evaluate a regime that current multimodal benchmarks (Luong et al., 2025b; Burgess et al., 2025; Yue et al., 2025; Li et al., 2024c,d) still underexpose: whether VLMs are truly reliable on elementary visual skills? As shown in **Fig. 5**, FBI-Bench occupies a distinct corner of the benchmark landscape. Its goal is not to be another broad-coverage benchmark (Yu et al., 2023; Chen et al., 2024b; Fu et al., 2025a) or another benchmark that pushes toward ever harder reasoning (Roberts et al., 2025; Lu et al., 2024; Luong et al., 2025b; Zhang et al., 2025b), but to ask a more basic question: *do strong results on existing multimodal benchmarks really mean that models have mastered the perceptual foundations those tasks presuppose?* This question motivates **FBI-Bench**. Rather than following the dominant trend toward scale and difficulty (Rein et al., 2023; Mialon et al., 2023), FBI-Bench is built to expose a different kind of weakness: simple, systematic failures that impressive headline performance can easily hide.

This focus follows directly from the motivation in **Appendix A.1**. Higher-level multimodal competence (Luong et al., 2025a,b) is only as reliable as the lower-level perceptual skills that support it. A frontier models (Singh et al., 2025; Gemini Team, 2025a; Bai et al., 2025a; Anthropic, 2025a) may perform well on complex reasoning benchmarks yet still fail on simple recognition, localization, relation, direction, counting, OCR, or comparison tasks. These failures matter because they are often the ones users notice first: in practice, such failures are especially important because they are often the ones users notice first: trust is more easily damaged by an obvious mistake on an easy case than by failure on an unusually difficult one (Burgess et al., 2025; Rein et al., 2023; Yue et al., 2024). From this perspective, FBI-Bench is intentionally centered on *fundamental* capabilities, not because these tasks are glamorous, but because they are the necessary substrate of reliable multimodal intelligence (Mumuni and Mumuni, 2025; Hendrycks et al., 2025).

This emphasis also explains why FBI-Bench is intentionally *impossible* for current models (OpenAI et al., 2024a; Gemini Team, 2025b; Google, 2025). As discussed in **Appendix A.2**, the goal is not rhetorical difficulty, but methodological diagnostic value. If a sample is failed by only one model, the error may simply reflect model-specific noise; but if the same human-easy sample is failed consistently across diverse frontier and open-weight models (Bai et al., 2025b; Team et al., 2025; Liu et al., 2023b; Team, 2025), it is much more likely to reveal a shared limitation of the field. This also helps FBI-Bench avoid rapid saturation. Many existing benchmarks either become less informative once models approach ceiling performance (Burgess et al., 2025; Rein et al., 2023; Masry et al., 2022), or move toward harder and more composite tasks, which can make foundational weaknesses harder to diagnose (Tang et al., 2025; Li et al., 2023). By contrast, FBI-Bench keeps the tasks simple while retaining only cases that remain unsolved across models, so that any future gain is easier to interpret as genuine progress in perceptual grounding. In this sense, its impossibility is not a gimmick, but an important way to preserve headroom while keeping the benchmark maximally diagnostic.

A further distinction lies in the evaluation protocol. Many multimodal benchmarks rely heavily on multiple-choice questions for efficiency and standardization (Chen et al., 2024b; Ying et al., 2024; Zhang et al., 2025b; Rein et al., 2023). However, as argued in **Appendix A.3**, MCQ evaluation can substantially overestimate model ability (Wang et al., 2024; Yao et al., 2025; Kanade and Ganu, 2025; Jiang et al., 2024), especially when the goal is to test foundational perception. A model may succeed by recognizing the least implausible option, exploiting answer-order bias, or relying on partial linguistic cues, without truly grounding its answer in the image (Ging et al., 2024; Zhang et al., 2025a; Atabuzzaman et al., 2025). This issue becomes particularly problematic for a benchmark like ours, where even slight option-induced assistance can blur the true source of apparent success (Yu et al., 2023; Liu et al., 2025b). FBI-Bench therefore adopts a free-form setting, forcing the model to produce what it actually perceives rather than select from candidate answers. This makes evaluation stricter, but also more accurate and faithful to the capability we aim to measure.

As discussed in **Appendix A.4**, **FBI-Bench** also revisits benchmark design from a different practical perspective: *diagnostic value does not necessarily scale with benchmark size*. Many re-

Table 7: Comparison of open-VLM benchmarks. QA Pairs denotes the number of question-answer pairs; Current SOTA denotes the best reported performance. Difficulty levels are abbreviated as EH (Extremely Hard), H (Hard), and M (Medium). Answer types are abbreviated as FF (Free-form), Y/N (Yes/NO Question) and MCQ (Multiple Choice Question).

Benchmark Name	QA Pairs	Current SOTA	Difficulty	Answer Type
VQAv2 (Goyal et al., 2017)	> 1M	80.80	M	FF
GQA (Hudson and Manning, 2019)	> 1M	67.10	M	FF
TextVQA (Singh et al., 2019)	45336	85.50	M	FF
QBench (Wu et al., 2024)	84773	80.50	M	Y/N + FF
QBench+ (Zhang et al., 2024b)	87222	81.07	M	Y/N + FF
MMBench (Liu et al., 2024d)	3217	82.00	M	MCQ
RealWorldQA (X.AI, 2024)	765	78.70	M	MCQ
ChartQA (Masry et al., 2022)	32719	90.80	M	FF
ScienceQA (Burgess et al., 2025)	21000	99.50	M	MCQ
VisIT-Bench (Bitton et al., 2023)	592	65.43	H	FF
BLINK (Fu et al., 2024)	3807	68.50	H	MCQ
SEEDBench (Ying et al., 2025)	19242	79.50	H	MCQ
MathVista (Lu et al., 2024)	735	82.80	H	MCQ + FF
MMBench (Liu et al., 2024c)	3217	88.90	H	FF
MMVet (Yu et al., 2023)	218	89.40	H	FF
MMStar (Chen et al., 2024b)	1500	82.10	H	MCQ
MMT-Bench (Ying et al., 2024)	32325	72.60	H	MCQ
MMMU (Yue et al., 2024)	11500	81.80	H	MCQ + FF
SEEDBench2+ (Li et al., 2024b)	2300	73.80	EH	MCQ
MME-RealWorld (Zhang et al., 2025b)	29429	61.60	EH	MCQ
MTVQA (Tang et al., 2025)	6778	36.80	EH	FF
GAIA (Mialon et al., 2023)	466	91.69	EH	FF
GPQA Diamond (Rein et al., 2023)	198	94.00	EH	MCQ
OCRBenchV2 (Fu et al., 2025c)	10000	63.40	EH	FF
ZeroBench (Roberts et al., 2025)	100	23.00	EH	FF
MMMU-Pro (Yue et al., 2025)	3460	82.00	EH	MCQ + FF
IMO Bench (Luong et al., 2025b)	400	65.70	EH	FF
FBI-Bench	**500**	**00.00**	**Fundamental**	**FF**

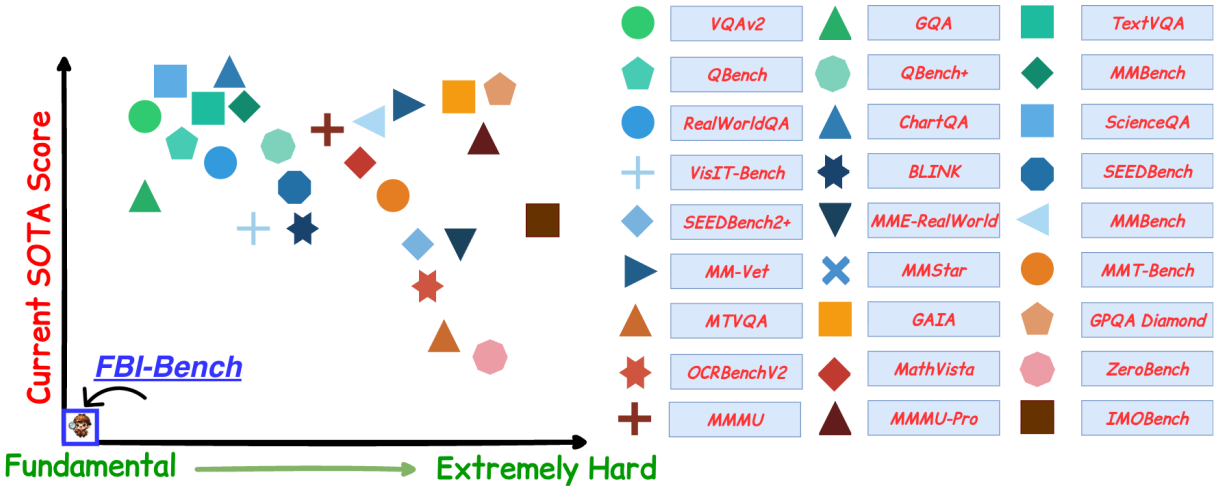


Figure 5: **Positioning of FBI-Bench in the landscape of open-VLM benchmarks.** Existing benchmarks are largely concentrated on medium-to-extremely hard tasks and are often associated with relatively high current SOTA scores. In contrast, FBI-Bench occupies a distinct corner: it targets fundamental, human-easy perceptual skills that remain systematically challenging for current models, revealing shared failures that may be obscured by strong performance on more difficult benchmarks.

cent multimodal benchmarks (Yao et al., 2025; Yue et al., 2025, 2024) contain tens of thousands or even millions of samples, but evaluating them across diverse frontier and open-source models – especially in free-form settings (Ying et al., 2025; Li et al., 2024b, 2023) – can be computationally expensive and difficult to repeat. More importantly, such scale often yields broad aggregate scores without necessarily revealing the shared, systematic failures that matter most for understanding model limitations. FBI-Bench therefore takes the opposite approach: it is intentionally compact and lightweight, yet each sample is highly information-dense, because it is human-easy, targets fundamental skills, and is retained only when it exposes a failure that persists across models. In this sense, the value of FBI-Bench lies not in raw scale, but in providing a more efficient and interpretable way to uncover systematic weaknesses in current VLMs.

These differences are reflected clearly in **Tab. 7**. If existing benchmarks tell us how far models can go on broad or difficult tasks, FBI-Bench tells us whether those gains rest on stable perceptual foundations in the first place. This is precisely why FBI-Bench appears separated in **Fig. 5** and distinct in **Tab. 7**: it is designed to reveal the basic, shared, and highly informative failures that other benchmarks can leave hidden. In our view, this is precisely the kind of benchmark needed to reveal systematic limitations, generate actionable insight, and make future progress more interpretable.

B.2 More About FBI-Bench

B.2.1 Ethical considerations

Although FBI-Bench is manually curated and carefully verified, it may still be subject to some degree of annotator-dependent variation or occasional imperfection. As described in Sec. 3.1, the benchmark is built through a multi-stage curation pipeline with repeated cleaning, annotator voting, and model-based filtering. These procedures are designed to improve data quality and consistency, but some borderline cases may still reflect differences in interpretation in sample selection, wording, or capability labeling. This is particularly relevant for a benchmark like ours, whose purpose is to identify human-easy examples that reveal shared weaknesses of current models, since such decisions inevitably involve a degree of human judgment even under careful quality control.

Accordingly, we view FBI-Bench as a carefully constructed diagnostic benchmark that is intended to complement, rather than replace, other evaluations of multimodal ability. More broadly, this is true of any benchmark that depends on human curation and qualitative decisions about what should count as *fundamental*, *unambiguous*. To make these decisions as reliable as possible, we employ diverse annotator recruitment, majority-based verification, repeated deduplication and cleaning, and human performance checks to ensure that the benchmark remains well-posed and interpretable. At the same time, we welcome feedback and con-

structive input from the community, which can further improve the benchmark over time and help refine individual cases where clarification may be beneficial.

B.2.2 Curation Funnel and Post-Filtering Validation

Stage	Curation Process	Samples	Retention
Initial	Raw image candidates and image-question candidates (X_0)	24,460	100%
Stage 1	Visual filtering (X_1)	18,285	74.8%
Stage 2	QA annotation, capability labeling, and human-easy verification (X_2)	10,142	41.5%
Stage 3a	Duplicate, ambiguity, irrelevance, and difficulty filtering (X_3)	5,318	21.7%
Stage 3b	**Model-based filtering and final sampling**	**500**	**9.4% of X_3**

Table 8: Three-stage curation funnel of **FBI-Bench**. Stage 1 filters visually suitable images, Stage 2 constructs and verifies question–answer annotations, and Stage 3 removes noisy or non-diagnostic samples before model-based filtering and final selection.

To rule out the possibility that the 0.0% accuracy floor is caused by noisy or ambiguous samples, we report the full curation funnel in **Tab. 8**. Starting from 24,460 raw candidates, Stage 1 collects images from public web sources and prior benchmarks, followed by visual filtering for image clarity, source suitability, and relevance to fundamental perception. Stage 2 converts the retained images into free-form question–answer samples: annotators write or refine questions, assign capability labels, and verify that each sample is human-easy and answerable from the image alone. Stage 3 removes duplicate, ambiguous, irrelevant, overly difficult, or incorrectly labeled samples through multi-round filtering and cross-verification. This process yields 5,318 high-quality candidates (X_3), from which the final 500 samples are selected through model-based rejection and sampling.

We further compare **FBI-Bench** with *FBI-Lite*, a 500-sample subset sampled from the pre-filtered candidate pool before model-based rejection, where at least one evaluated model succeeds. As shown in **Tab. 2**, both sets exhibit similar distributions of low-level visual confounders such as small objects, occlusion, and high-density scenes. This indicates that the model collapse on **FBI-Bench** is unlikely to be an artifact of adversarial image quality or unusually poor visual conditions. Instead, the contrast between **FBI-Bench** and FBI-Lite suggests that many apparent model successes in the pre-filtered pool may rely on canonical biases or visual shortcuts, while **FBI-Bench** isolates cases that require more reliable fine-grained visual grounding.

B.2.3 A benchmark is most useful when it becomes obsolete

More broadly, we do not view FBI-Bench as a benchmark that should remain unsolved for as long as possible. In fact, we hope the opposite. Because FBI-Bench targets elementary capabilities that should be routine for reliable multimodal systems, its long-term value lies not in staying permanently difficult, but in making current weaknesses visible enough that future models can overcome them. A benchmark centered on basic perception should ideally be solved quickly once the field develops the right representations, alignment strategies, and evaluation practices. If FBI-Bench is solved in the near future, that would not diminish its contribution; rather, it would be strong evidence that the community has made real progress on the perceptual foundations that current systems still lack.

This perspective also reflects the central philosophy of the benchmark. FBI-Bench is not designed to create artificial difficulty for its own sake. Instead, it is designed to expose a gap that existing evaluations can overlook: models may perform strongly on prestigious or complex benchmarks while still failing on simple visual judgments that humans handle effortlessly. In that sense, the ideal outcome is for FBI-Bench to be solved *as soon as possible*, because such a result would indicate that models are no longer merely impressive on difficult demonstrations, but are also becoming dependable on the basic skills that more advanced multimodal reasoning presupposes.

B.2.4 Release and community use

To support transparency, reproducibility, and follow-up research, we plan to publicly release the benchmark and the corresponding evaluation code upon publication of the paper. We hope this release will make FBI-Bench easy to inspect, reproduce, stress-test, and extend. Public availability is particularly important for a benchmark of this kind, since its utility depends not only on reported headline numbers, but also on whether other researchers can examine individual cases, analyze failure patterns, challenge questionable samples, and build improved methods directly against the benchmark.

We also hope the release will encourage the community to treat FBI-Bench not only as a leaderboard-style test, but as a shared diagnostic resource. Its purpose is to help identify where current VLMs still fail at the foundations, and, just as im-

portantly, to provide a concrete target for methods that aim to strengthen those foundations. In this sense, we see FBI-Bench as an open benchmark in both senses of the term: open in availability, and open to refinement through community scrutiny and future iteration.

B.3 Future Work

Our findings open several directions for future research.

- **FBI-Bench should be used not only as an evaluation set, but also as a diagnostic instrument for uncovering systematic limitations in current vision-language models.** A key design choice of FBI-Bench is that benchmark samples are retained only when they are failed consistently across diverse models, which helps reduce model-specific noise and makes the remaining errors more likely to reflect shared weaknesses rather than isolated implementation artifacts. Combined with explicit annotations over seven foundational skills and their compositions, this structure creates an opportunity for substantially deeper analysis of where failures originate: whether from the visual encoder, multimodal alignment, post-training, or the interaction between perception and reasoning. In this sense, a natural next step is to move beyond aggregate accuracy and build finer-grained studies of error taxonomy, cross-skill dependency, failure transfer across architectures, and representation-level diagnostics that can reveal the root causes of persistent perceptual errors.

- **Developing *methods* that explicitly improve the foundational capabilities measured by FBI-Bench.** Our experiments suggest that many commonly adopted strategies, including visual hinting, latent reasoning, standard post-training, additional pretraining, and scaling the language backbone, provide only limited or unstable gains on these tasks, whereas improvements in vision encoder design and data alignment appear to matter more. This implies that fundamental perception is unlikely to emerge automatically from more reasoning compute or larger language models alone. Instead, progress may require methods that directly strengthen spatial and structural grounding at the representation level, preserve perceptual precision during alignment, and improve coordination across composed capabilities such as localization with relation reasoning or counting with OCR. This direction is well motivated by both our findings and prior diagnostic work showing that models can succeed on difficult multimodal tasks while still relying on shortcuts or shallow cues in perception-heavy settings. More broadly, improving FBI-Bench-style capabilities is not only valuable for this benchmark itself: because elementary perception underpins reliable multimodal reasoning, gains on these foundational skills may be a necessary condition for more robust performance on complex downstream tasks.

Overall, we view FBI-Bench as a starting point rather than an endpoint. Its main value lies in exposing a gap that current evaluation protocols and scaling trends can easily overlook: strong performance on complex benchmarks does not necessarily imply mastery of the perceptual foundations that those benchmarks presuppose. Future work should therefore pursue the problem on three fronts simultaneously: deeper causal analysis of shared failures, broader and more comprehensive benchmark construction, and model designs that explicitly target the perceptual bottlenecks revealed by FBI-Bench.

C Experimental Setting

C.1 Prompts

To avoid prompting artifacts that may artificially improve reasoning performance, we enforce a strict direct-answer format for all models.

Standardized Prompt Template

```
{image(s)}
{question}
Only provide the answer; no reasoning steps
in the answer.
If you are unsure, still provide an answer.
Answer:
```

Due to the open-ended nature of free-form answers, we employ an LLM-based judge to evaluate whether model predictions semantically match the reference answers.

LLM-Based Evaluator Prompt

Role
`You are an AI Judge. Compare the Model Prediction with the Ground Truth.`

Context
`Question: "{question}"`
`Ground Truth: "{gt}"`
`Model Prediction: "{pred}"`

Task

- Determine how well the `Model Prediction` matches the `Ground Truth` semantically.
- Ignore differences in case, punctuation, or minor phrasing.
- Score correctness on a continuous scale: 0.0 (totally wrong) to 1.0 (totally right), using increments of 0.1 only. Example valid scores: 0.0, 0.1, 0.2, . . . , 0.9, 1.0.
- Choose the closest score that reflects the semantic match quality.
- `<AND>` requires all components to be present in the prediction.
- `<OR>` requires any one valid component to be present.

Output Format
`{"score": 0.0 // one of 0.0, 0.1, . . . , 0.9, 1.0}`

To construct multiple-choice variants of the questions, we additionally use an LLM to generate plausible distractor answers grounded in the visual content.

MCQ Distractor Generation Prompt

`You are creating multiple-choice distractors for a visual question answering benchmark.`
`Look at the image carefully, then consider:`
`Question: "{question}"`
`Correct Answer: "{answer}"`
`Generate exactly 3 plausible but INCORRECT answer choices for this question.`
`The wrong answers must:`

- Be grounded in what is visually present in the image.
- Be clearly wrong compared to the correct answer.
- Be concise (similar length/style to the correct answer).

Output JSON Only

```
{
"wrong_answer_1": "...",
"wrong_answer_2": "...",
"wrong_answer_3": "..."
}
```

C.2 Evaluation Protocol and Meta-Evaluation

Open-ended evaluation. FBI-Bench adopts an open-ended generation protocol instead of a Multiple-Choice Question (MCQ) format. This design avoids answer-option shortcuts and better reflects whether a model can directly produce the required visual attribute, such as an object category, count, color, spatial relation, direction, or OCR content. Each model prediction is anonymized before evaluation.

Meta-evaluation of the LLM judge. To select a reliable automated evaluator, we conduct a meta-evaluation against human annotations. We randomly sample $n = 100$ prediction–answer pairs and obtain binary human labels indicating whether each prediction is correct. We then compare these human labels with three automated judge configurations: Gemini-only, GPT-only, and a GPT+Gemini ensemble. For each configuration, we compute the average absolute discrepancy:

$$\bar{\Delta} = \frac{1}{N}\sum_{i=1}^{N} |s_i^{\text{judge}} - s_i^{\text{human}}|,$$

where $s_i^{\text{human}} \in \{0, 1\}$ is the human score and $s_i^{\text{judge}} \in [0, 1]$ is the automated score. A lower $\bar{\Delta}$ indicates stronger agreement with human judgment. The configuration with the lowest discrepancy is selected as the primary evaluator for the main experiments.

Metric	Gemini Only	GPT Only	Ensemble (GPT + Gemini)
$\bar{\Delta}$ (↓)	0.063	0.085	0.051

Table 9: Average absolute discrepancy $\bar{\Delta}$ between automated LLM evaluators and human annotations on a random subset of 100 prediction–answer pairs. Lower values indicate stronger agreement with human judgment.

Full benchmark evaluation. For the full benchmark, the selected judge evaluates whether each model prediction is semantically equivalent to the reference answer. The judge is instructed to ignore minor surface variations, including casing, punctuation, formatting, and synonymous expressions, while focusing strictly on the correctness of the required atomic visual attributes. Verbose explanations are not rewarded unless they contain the correct answer. Each prediction receives a continuous score from 0.0 to 1.0 in increments of 0.1, and we classify scores ≥ 0.5 as correct.

Stability check. To assess evaluation stability, we repeat the automated evaluation multiple times under the same rubric and report the consistency of judge decisions. In addition, we manually inspect representative predictions to verify that the automated judgments reflect the intended correctness criteria.

C.3 Hyperparameters

The decoding hyperparameters are summarized in Tab. 10.

C.4 Models

To ensure full reproducibility, we report the exact model versions or checkpoints used in our experiments. For proprietary models accessed through APIs, we provide the official API identifiers. For open-source models, we report the exact HuggingFace repository names or checkpoint identifiers whenever available.

Proprietary Models

- GPT-4o: `gpt-4o`
- GPT-5 Mini: `gpt-5-mini`
- Claude 4.5 Haiku: `claude-haiku-4-5-20251001`
- Claude 4.5 Sonnet: `claude-sonnet-4-5-20250929`
- Gemini 2.5 Flash: `gemini-2.5-flash`
- Gemini 2.5 Pro: `gemini-2.5-pro`
- Gemini 2.0 Flash: `gemini-2.0-flash`

Open-weighted Vision–Language Models

- InternVL2 8B: `OpenGVLab/InternVL2-8B`
- InternVL2.5 78B: `OpenGVLab/InternVL2_5-78B`
- InternVL3 78B: `OpenGVLab/InternVL3-78B`
- InternVL3.5 1B: `OpenGVLab/InternVL3_5-1B-HF`
- InternVL3.5 8B: `OpenGVLab/InternVL3_5-8B`
- InternVL3.5 14B: `OpenGVLab/InternVL3_5-14B-Instruct`
- InternVL3.5 38B: `OpenGVLab/InternVL3_5-38B`
- Qwen2-VL 7B: `Qwen/Qwen2-VL-7B`
- Qwen2-VL 7B Instruct: `Qwen/Qwen2-VL-7B-Instruct`
- Qwen2.5-VL 72B Instruct: `Qwen/Qwen2.5-VL-72B-Instruct`
- Qwen3-VL 2B Instruct: `Qwen/Qwen3-VL-2B-Instruct`
- Qwen3-VL 8B Instruct: `Qwen/Qwen3-VL-8B-Instruct`
- Kimi-VL A3B Instruct: `moonshotai/Kimi-VL-A3B-Instruct`
- Kimi-VL A3B Thinking: `moonshotai/Kimi-VL-A3B-Thinking`
- LLaVA 1.5 7B: `liuhaotian/llava-v1.5-7b`
- LLaVA 1.5 13B: `liuhaotian/llava-v1.5-13b`
- LLaVA v1.6 Vicuna 7B: `llava-hf/llava-v1.6-vicuna-7b-hf`
- LLaVA v1.6 Mistral 7B: `llava-hf/llava-v1.6-mistral-7b-hf`

Table 10: **Hyperparameters used in model decoding setting**.

Hyperparameter	Greedy Decoding (Main Evaluation)	Stochastic Decoding (Consistency Analysis)
Temperature	0.0	0.7
Top-k	1	–
Top-p	–	0.95
Seed	42	42
Number of Runs	1	3

- LLaVA-OneVision 0.5B: `llava-hf/llava-onevision-qwen2-0.5b-ov-hf`
- LLaVA-OneVision 7B: `llava-hf/llava-onevision-qwen2-7b-ov-hf`
- LLaVA-OneVision 72B: `llava-hf/llava-onevision-qwen2-72b-ov-hf`
- Gemma 3 1B Instruct: `google/gemma-3-1b-it`
- Gemma 3 4B Instruct: `google/gemma-3-4b-it`
- Gemma 3 12B Instruct: `google/gemma-3-12b-it`
- Gemma 3 27B Instruct: `google/gemma-3-27b-it`
- PaliGemma 3B PT 224: `google/paligemma-3b-pt-224`
- PaliGemma 3B PT 448: `google/paligemma-3b-pt-448`
- PaliGemma 3B PT 896: `google/paligemma-3b-pt-896`
- 360VL 70B: `qihoo360/360VL-70B`
- IDEFICS2 8B: `HuggingFaceM4/idefics2-8b`
- Molmo 72B: `allenai/Molmo-72B-0924`

C.5 Human Baseline

To verify that FBI-Bench is genuinely solvable by humans under reasonable evaluation conditions, we conducted a blind human baseline study on both FBI-Core and FBI-Lite. The study involved 30 participants from three education levels: 10 primary-school students, 10 middle-school students, and 10 high-school students. None of the participants were involved in dataset construction, question writing, answer annotation, or model-output verification.

Protocol. Participants answered the same open-ended visual questions used in the VLM evaluation. Each trial contained only the image and the corresponding free-form question, without multiple-choice options, visual hints, bounding boxes, or additional context. Participants were instructed to provide a short direct answer and were not asked to explain their reasoning. External tools, search engines, calculators, and communication with other participants were not allowed. The evaluation was conducted in a blind setting: participants did not know whether a sample came from FBI-Core or FBI-Lite, and they were not shown model predictions or ground-truth answers during annotation. Questions were presented in randomized order to reduce ordering effects. Participants were allowed to inspect the image carefully, but no information beyond the image and the question was provided.

Scoring. Human answers were scored against the same canonical answers and aliases used for model evaluation. An answer was marked correct if it semantically matched the ground truth, allowing minor variations in wording, spelling, synonym usage, or equivalent numerical expressions. For questions requiring multiple visual components, all required components had to be correct. Ambiguous, incomplete, contradictory, or uncertain answers were marked as incorrect. The scoring criterion focused only on the target visual attribute, such as object identity, location, spatial relation, direction, count, OCR text, or comparison result, and did not reward additional explanation.

Human verification of model failures. In addition to measuring human accuracy, we manually verified the model-failure cases used in FBI-Core. This check ensures that the observed model collapse is not caused by ambiguous ground truth, unreasonable questions, or intrinsically unsolvable samples. The 0.0% floor reported for the main VLM evaluation was human-verified, and

Table 11: **Human baseline on FBI-Bench.** Accuracy is reported in %. Participants answer the same open-ended visual questions used for VLM evaluation.

Group	#Participants	Location	Relation	Recognition	Counting	Direction	OCR	Comparison	Overall
Primary School	10	89.3	86.7	81.5	77.4	73.2	69.1	61.8	81.2
Middle School	10	94.1	92.6	87.4	84.8	82.9	79.6	76.3	87.8
High School	10	98.2	97.6	96.8	95.9	94.7	93.8	91.5	96.4

our judge-based evaluation was further checked against human consensus on FBI-Lite, where the judge score closely matched the human-verified result. This supports the reliability of the automatic evaluator while confirming that the retained FBI-Core samples are human-solvable but systematically difficult for current VLMs.

Results. Tab. 11 reports the human baseline by education level and skill category. Primary-school students achieve 81.2% overall accuracy, while middle-school and high-school students achieve 87.8% and 96.4%, respectively. The per-skill breakdown further shows that humans perform strongly across all seven fundamental capabilities, including localization, relation understanding, recognition, counting, direction, OCR, and comparison. These results confirm that FBI-Bench is not an adversarial or impossible benchmark for humans. Instead, the large human–model gap indicates that FBI-Bench exposes model-specific failures in fundamental visual perception.

D Additional Ablation Studies

D.1 Ablation 1: The Zero-Consistency Dilemma and Visual Lottery

To thoroughly assess the robustness of MLLMs on fundamental visual tasks, we analyze their prediction consistency by contrasting top-3 and top-1 strict matching accuracy. The results, detailed in **Tab.12**, expose a catastrophic failure in confidence calibration across all state-of-the-art models.

While top-3 evaluation yields marginal accuracy (e.g., 1.61% overall for Gemini 2.5 Flash(Gemini Team, 2025b), with scattered successes like 1.96% on Location and 1.92% on Relation), an alarming and definitive pattern emerges under top-1 evaluation: **Every single model, regardless of size or proprietary status, drops to exactly 0.0%**.

This zero-consistency phenomenon implies that MLLMs are essentially treating visual grounding as a "visual lottery". They are capable of generating a distribution of plausible linguistic guesses that occasionally encompass the correct structural answer within the top 3 candidates. However, the actual visual signal connecting the image pixels to the semantic tokens is too weak to spike the logits of the correct answer to rank 1. In other words, when models happen to be correct, it is a byproduct of stochastic text generation rather than genuine visual certainty.

D.2 Ablation 2: Open-ended Evaluation vs. The Multiple-Choice Illusion

To directly verify the claims in **Sec. A.3**, we compare MCQ and free-form evaluation on **100 randomly sampled FBI-Bench instances** across five models spanning a range of architectures and scales (**Tab. 13**).

All models improve under MCQ. Consistent with the claim in **Sec. A.3**, *every* model benefits from the MCQ format. Models that score 0.0% under strict free-form evaluation (Qwen3-VL-8B-Instruct(Bai et al., 2025a), Kimi-VL-A3B-Instruct(Team et al., 2025)) still achieve 21% and 30% respectively under MCQ on the same 100 instances—gains that cannot reflect genuine visual understanding and must arise from option-recognition and guessing artifacts.

MCQ advantage amplifies with scale. The InternVL3.5(Zhu et al., 2025) family provides the clearest evidence for the scale-dependent MCQ inflation discussed in **Sec. A.3**. Under free-form evaluation, performance degrades with scale (8B: 3.4% → 78B: 1.3%), a signature of semantic overshadowing (**Ablation D.5**). Yet under MCQ on this 100-sample subset, the 78B model jumps to 55%—more than double the 8B score of 23%. This stark reversal directly supports the claim that scaling amplifies the ability to exploit option artifacts rather than genuine visual grounding: a stronger language backbone is better at selecting the least-wrong candidate, even when its perceptual representations remain structurally deficient.

D.3 Ablation 3: Can Prompt Rethinking Trigger Visual Self-Correction? (Supplementary)

Building upon Ablation 2 in the main paper, we provide additional extensive results to investigate whether explicitly prompting models to "recheck" their answers (a pseudo-Chain-of-Thought approach) can trigger visual self-correction. **Tab. 14** reveals a counter-intuitive phenomenon: rather than correcting errors, rethinking often exacerbates them through anchoring bias.

For standard instruction-tuned models (e.g., Qwen3-VL-8B-Inst(Bai et al., 2025a)), the recheck prompt artificially inflates scores from 0.0% to 5.0%, merely restoring performance back to the baseline level of native Thinking models. However, the most critical insight comes from observing native Thinking models on tasks they are relatively proficient at. For instance, on the {Cou, OCR} combination, Qwen3-VL-8B-Thinking(Bai et al., 2025a) plummets from 43.8% to 25.0% when forced to re-evaluate, and completely fails (21.4% → 0.0%) on {Dir, Loc, Rec}.

This reveals a profound disconnection between the perception and reasoning modules. When prompted to rethink, the language module does not command the vision encoder to perform a fresh, higher-resolution spatial scan. Instead, the model falls victim to anchoring bias: it utilizes the additional generation steps to confidently confabulate logical justifications defending its flawed initial visual premise. Increasing reasoning compute cannot rectify fundamental perceptual errors if the initial visual tokens are structurally malformed.

D.4 Ablation 4: The Alignment Tax on Visual Grounding (Supplementary)

To provide deeper context for Ablation 3 in the main paper, we investigate the pre-trained base and instruction-tuned variants of the Qwen2-VL(Bai

Table 12: Performance on FBI-Bench. Accuracy is reported in % using top-3 predictions. All models obtain exactly 0% under top-1 evaluation, indicating that none can correctly solve the benchmark under strict matching.

Model	Location	Relation	Recognition	Counting	Direction	OCR	Comparison	Overall
Proprietary MLLMs								
GPT-4o(OpenAI et al., 2024a)	1.72	1.84	1.63	1.51	0.92	1.96	0.71	1.47
GPT-5-mini(Singh et al., 2025)	1.21	1.43	1.32	1.02	0.41	0.36	0.71	1.05
Claude Haiku 4.5(Anthropic, 2025b)	0.92	0.88	0.94	0.31	0.21	0.63	1.58	0.78
Claude Sonnet 4.5(Anthropic, 2025a)	0.63	0.59	0.52	0.19	0.41	0.00	1.84	0.61
Gemini 2.5 Flash(Gemini Team, 2025b)	**1.96**	**1.92**	**1.71**	**1.63**	**1.52**	1.84	0.71	**1.61**
Gemini 2.0 Flash(Google, 2025)	0.52	0.49	0.84	0.63	0.21	0.00	0.71	0.49
Open-weight MLLMs								
360VL-70B(Qihoo 360 and 360CVGroup, 2024)	0.52	0.36	0.84	0.92	0.41	0.63	0.00	0.52
InternVL2.5-78B(Chen et al., 2025b)	0.36	0.31	0.71	0.31	0.21	0.36	0.00	0.32
InternVL3-78B(Zhu et al., 2025)	0.41	0.36	0.63	0.00	0.21	0.00	1.12	0.39
Kimi-VL-A3B-Instruct(Team et al., 2025)	0.71	0.92	0.84	0.19	0.00	0.00	1.41	0.63
Qwen2.5-VL-72B-Instruct(Bai et al., 2025b)	0.21	0.16	0.31	0.31	0.41	0.63	0.00	0.29
Qwen3-VL-8B-Instruct(Bai et al., 2025a)	**0.92**	**0.71**	**1.32**	**1.41**	**0.41**	**1.63**	0.00	**0.91**
Molmo-72B(Deitke et al., 2024)	0.41	0.39	0.16	0.16	0.63	0.00	0.00	0.29
LLaVA-OneVision-Qwen2-7B(Li et al., 2024a)	0.52	0.31	0.84	0.31	0.41	0.63	0.00	0.43

Table 13: Ablation 2: MCQ vs. free-form evaluation on 100 randomly sampled FBI-Bench instances. Accuracy is presented in %.

Model	**Params**	**MCQ Accuracy on 100 random samples (%)**						
		Location	Relation	Recognition	Counting	Direction	OCR	Overall
Kimi-VL-A3B-Instruct(Team et al., 2025)	3B	29.5	31.4	26.1	29.2	21.4	40.0	30
Qwen3-VL-8B-Instruct(Bai et al., 2025a)	8B	20.5	21.6	21.7	8.3	14.3	6.7	21
LLaVA-OneVision-1.5-8B(Li et al., 2024a)	8B	19.2	21.6	19.6	16.7	7.1	20.0	21
InternVL3.5-8B(Zhu et al., 2025)	8B	23.1	23.5	19.6	16.7	14.3	20.0	23
InternVL3.5-78B(Zhu et al., 2025)	78B	61.5	68.6	43.5	25.0	64.3	40.0	**55**

Table 14: Ablation 3 (Supplementary): Evaluation results on MLLMs with or not latent thinking regarding each capability integration. Accuracy is presented in %.

Model	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Qwen3-VL-8B-Inst.(Bai et al., 2025a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-8B-Inst. w/ recheck(Bai et al., 2025a)	5.0	3.9	5.3	4.8	16.0	0.0	4.3	6.3	0.0	6.3	14.3	0.0	16.7	5.0
Qwen3-VL-8B-Thinking(Bai et al., 2025a)	5.0	2.0	1.8	4.8	12.0	8.3	8.7	0.0	43.8	6.3	21.4	7.7	8.3	6.0
Qwen3-VL-8B-Thinking w/ recheck(Bai et al., 2025a)	7.1	4.9	5.3	7.1	4.0	0.0	4.3	6.3	25.0	6.3	0.0	15.4	8.3	6.0
Kimi-VL-A3B-Instruct(Team et al., 2025)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kimi-VL-A3B-Instruct w/ recheck(Team et al., 2025)	8.57	4.9	0	0	0	8.3	0	6.3	0	0	0	0	8.3	4.0
InternVL3.5-8B(Zhu et al., 2025)	2.1	1.9	8.8	0	12.0	4.1	4.3	12.5	0	0	0	0	0	3.0
InternVL3.5-8B w/ recheck(Zhu et al., 2025)	3.6	2.9	7.0	7.1	12.0	0	8.7	0	0	6.3	0	0	0	4.0

et al., 2025b) and LLaVA-OneVision(Li et al., 2024a) families (**Tab. 15**).

The results consistently expose what we term the "Alignment Tax". Across both families, Base models retain better raw perceptual metrics than their Instruct counterparts. For example, after instruction tuning, Qwen2-VL-7B(Bai et al., 2025b) suffers notable degradations in highly complex structural combinations: dropping from 10.8% to 6.9% on {Loc, Rec, Rel}, and halving its performance on {Cou, Loc, OCR, Rec} (from 16.0% to 8.0%).

This phenomenon indicates that standard Supervised Fine-Tuning (SFT)(von Werra et al., 2020) emphasizes conversational fluency, formatting, and safety at the direct expense of visual precision. The optimization process forces the model to heavily weight its language priors to generate pleasing, human-like responses, thereby suppressing the noisy but fundamentally crucial fine-grained spatial representations learned during contrastive pre-training.

D.5 Ablation 5: Inverse Scaling and Semantic Overshadowing (Supplementary)

A prevailing assumption in the era of Large Language Models is the "Scaling Law," positing that increasing model parameters monotonically improves performance. To verify if this holds true for fundamental visual perception, we evaluate model families across varying parameter scales in **Tab. 16**.

The data strikingly contradicts standard scaling laws, revealing a clear pattern of **inverse scaling** on raw perceptual tasks. As parameter counts increase, performance consistently drops: InternVL3.5(Zhu et al., 2025) plunges from 11.0% (1B) to 2.2% (38B), LLaVA-OneVision(Li et al., 2024a) drops from 9.8% (0.5B) to an abysmal 0.0% (72B), and Qwen3-VL(Bai et al., 2025a) regresses from 7.8% (2B) to 4.4% (32B). (Gemma(Team, 2025) remains a rare exception).

We attribute this to Semantic Overshadowing. Larger LLM backbones possess massive parametric memories and overwhelmingly strong language priors. When faced with ambiguous or challenging visual inputs—such as counting overlapping objects or deducing fine-grained spatial directions—smaller models are forced to rely on the raw visual tokens, occasionally succeeding. Conversely, gargantuan 72B language models tend to ignore the weak visual signals entirely, opting instead to auto-regressively hallucinate a highly confident, linguistically coherent, but factually incorrect answer. Scaling up the language brain without fundamentally overhauling the vision encoder's structural capacity merely creates a more articulate confabulator.

D.6 Ablation 6: Decoding Settings and Representation Collapse

To unequivocally prove that the observed catastrophic failures are architectural rather than artifacts of sub-optimal text generation, we evaluate InternVL3.5-8B(Zhu et al., 2025) across a broad spectrum of decoding hyperparameters (Temperature, Top-p, and Beam Search).

Tab. 17 shows that changing decoding parameters has little effect: overall accuracy stays between 3.2% and 4.4%. Increasing the beam width to 5 even reduces performance from 4.4% to 3.2%, while Cou, Rec drop from 10.5% to 7.0%. This suggests a representation collapse: the model fails to encode the correct spatial information in its internal representation, so different decoding strategies cannot recover the correct answer.

D.7 Ablation 7: Resolution Scaling and Selective Spatial Gain

We evaluate **PaliGemma-3B**(Beyer* et al., 2024) across three native pre-training resolutions (224, 448, 896) to test whether finer pixel grids can alleviate visual grounding failures. Results are shown in **Tab. 18**.

Overall accuracy improves monotonically from 19.6% to 24.0%, yet the gains are highly uneven. Higher resolution selectively benefits spatial-relational combinations: {Loc, Rel} gains +7.2pp and {Dir, Loc, Rel} surges from 7.7% to 30.8% (+23.1pp) at 896, as denser pixel grids sharpen object boundary perception. In contrast, combinations involving Counting and OCR (e.g., {Cou, Loc, OCR}, {Cou, OCR}) remain frozen at 0.0% across all resolutions, exposing a structural ceiling independent of input fidelity. Notably, an intermediate resolution (448) triggers non-monotonic degradation on {Loc, Rec} (16.7% $\rightarrow$ 0.0%) and {Dir, Loc, Rec} (14.3% $\rightarrow$ 0.0%), suggesting that mid-range resolutions disrupt learned tokenization patterns without yet compensating with sufficient spatial density. These findings confirm that resolution scaling is a partial amplifier for spatial tasks, not a remedy for the fundamental perceptual failures observed on **FBI-Bench**.

Table 15: Ablation 4 (Supplementary): Evaluation results on Base vs. Instruct MLLMs regarding each capability integration. Accuracy is presented in %.

Model	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
LLaVA-OneVision-1.5-8B-Base(Li et al., 2024a)	5.0	4.9	0.0	2.4	0.0	16.7	13.0	0.0	0.0	6.3	0.0	0.0	0.0	4.0
LLaVA-OneVision-1.5-8B-Instruct(Li et al., 2024a)	2.1	2.0	1.8	2.4	4.0	16.7	4.3	0.0	0.0	0.0	0.0	0.0	0.0	3.0
Qwen2-VL-7B-Base(Bai et al., 2025b)	2.9	**10.8**	3.5	4.8	**16.0**	4.2	8.7	0.0	0.0	12.5	0.0	**15.4**	0.0	6.0
Qwen2-VL-7B-Instruct(Bai et al., 2025b)	3.6	6.9	0.0	4.8	8.0	16.7	4.3	0.0	0.0	6.3	7.1	0.0	8.3	5.0

Table 16: Ablation 5 (Supplementary): Detailed Param Scaling results across multiple capability combinations. Accuracy is presented in %.

Model	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Llava-1.5-7b(Liu et al., 2023a)	3.6	7.8	1.8	11.9	0.0	4.2	13.0	0.0	0.0	37.5	0.0	0.0	0.0	5.8
Llava-1.5-13b(Liu et al., 2023a)	3.6	2.9	0.0	4.8	4.0	29.2	4.4	0.0	0.0	12.5	0.0	0.0	0.0	4.2
Qwen3-VL-2B-Instruct(Bai et al., 2025a)	11.4	9.8	1.8	0.0	0.0	20.8	4.4	0.0	0.0	25.0	7.1	7.7	0.0	**7.8**
Qwen3-VL-4B-Instruct(Bai et al., 2025a)	2.1	2.0	3.5	0.0	8.0	0.0	17.4	6.2	0.0	0.0	14.3	0.0	0.0	3.2
Qwen3-VL-8B-Instruct(Bai et al., 2025a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-32B-Instruct(Bai et al., 2025a)	2.1	2.9	5.3	4.8	4.0	12.5	4.4	25.0	0.0	0.0	0.0	0.0	16.7	4.4
llava-onevision-qwen2-0.5b-si(Li et al., 2024a)	13.6	12.8	1.8	4.8	12.0	25.0	4.4	6.2	0.0	6.2	7.1	0.0	8.3	**9.8**
llava-onevision-qwen2-7b-si(Li et al., 2024a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
llava-onevision-qwen2-72b-si(Li et al., 2024a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
InternVL3.5-1B-HF(Zhu et al., 2025)	7.9	12.8	10.5	9.5	4.0	33.3	13.0	18.8	0.0	18.8	7.1	15.4	0.0	**11.0**
InternVL3.5-8B-HF(Zhu et al., 2025)	2.1	2.0	8.8	0.0	12.0	4.2	4.4	12.5	0.0	0.0	0.0	0.0	0.0	3.4
InternVL3.5-38B-HF(Zhu et al., 2025)	2.9	2.9	0.0	2.4	0.0	4.2	8.7	0.0	0.0	0.0	0.0	0.0	0.0	2.2
gemma-3-1b-it(Team, 2025)	12.1	16.7	10.5	0.0	8.0	20.8	0.0	0.0	0.0	6.2	0.0	7.7	8.3	10.0
gemma-3-27b-it(Team, 2025)	18.6	24.5	8.8	2.4	0.0	16.7	4.4	6.2	0.0	18.8	0.0	7.7	0.0	**13.4**

Table 17: Ablation 6: Evaluation results on InternVL3.5-8B-HF across various decoding settings. Accuracy is presented in %.

Model	Temp	Top-p	Beam	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
InternVL3.5-8B(Zhu et al., 2025)	0.0	1.0	1	1.4	2.9	10.5	2.4	12.0	8.3	4.4	12.5	0.0	0.0	7.1	0.0	8.3	4.4
InternVL3.5-8B(Zhu et al., 2025)	0.0	1.0	**5**	1.4	3.9	**7.0**	0.0	12.0	0.0	4.4	12.5	0.0	0.0	0.0	0.0	0.0	**3.2**
InternVL3.5-8B(Zhu et al., 2025)	0.5	1.0	1	1.4	2.9	12.3	0.0	8.0	0.0	4.4	12.5	0.0	6.2	7.1	0.0	0.0	3.8
InternVL3.5-8B(Zhu et al., 2025)	0.7	0.9	1	2.1	2.9	12.3	0.0	12.0	0.0	8.7	12.5	0.0	6.2	0.0	0.0	0.0	4.2
InternVL3.5-8B(Zhu et al., 2025)	1.0	1.0	1	1.4	2.9	12.3	0.0	8.0	0.0	8.7	12.5	0.0	6.2	0.0	0.0	0.0	3.8

Table 18: Ablation 7: Evaluation results on PaliGemma-3B across varying input resolutions. Accuracy is presented in %.

Model	Res.	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
PaliGemma-3B(Beyer* et al., 2024)	224	25.7	30.4	3.5	11.9	0.0	41.7	8.7	0.0	0.0	43.8	14.3	7.7	16.7	19.6
PaliGemma-3B(Beyer* et al., 2024)	448	27.9	33.3	1.8	19.1	4.0	45.8	13.0	0.0	0.0	43.8	0.0	7.7	0.0	21.0
PaliGemma-3B(Beyer* et al., 2024)	896	**32.9**	**39.2**	1.8	16.7	4.0	41.7	8.7	0.0	0.0	43.8	14.3	**30.8**	0.0	**24.0**

D.8 Ablaction 8: Does data pretraining affect the fundamental skills of VLMs?

To answer this, we evaluated LLaVA-v1.5-13B(Liu et al., 2023a) before and after pre-trained with SpatialVLM(Chen et al., 2024a)) and MS COCO Captioning(Chen et al., 2015) dataset(Han et al., 2024)) on the **Tab. 19**.

Overall performance remains extremely low across all settings, indicating that additional pretraining does not meaningfully improve fundamental perception. Gains are also inconsistent across capability combinations. Caption-based pretraining improves {Loc, Rel} from 3.6% to 8.6% and {Loc, Rec, Rel} from 2.9% to 8.8%, but yields no improvement for counting-related tasks, which remain at 0.0%. Compared with SpatialVLM, caption-based pretraining shows slightly stronger gains on localization-related skills, yet neither strategy consistently improves performance across capability compositions.

These results suggest that the benefits of additional pretraining are limited and unstable. Even datasets explicitly designed to enhance spatial reasoning fail to reliably improve tasks requiring spatial grounding, indicating that the bottleneck likely lies in the underlying visual representations rather than the pretraining data itself.

> ***Finding 4.*** Spatial awareness is an architectural deficit, not a data-scarcity problem. Flooding the model with coordinate annotations cannot compensate for a vision encoder that is fundamentally incapable of structural grounding.

D.9 Ablation 9: How do different post-training alignment strategies impact the preservation and development of fundamental visual skills?

We study how standard alignment procedures affect fundamental visual skills by comparing pre-trained, SFT(von Werra et al., 2020), and RL-based (RLHF/DPO) variants(Sun et al., 2023).

Tab. 20 shows that supervised fine-tuning (SFT) significantly degrades raw spatial perception: for the 7B model, overall accuracy drops from 6.0% to 2.0%. RL-based optimization (RLHF/DPO) performs better, reaching 9.0% accuracy, suggesting it better preserves useful internal representations that SFT tends to suppress. However, the gains remain modest. Notably, all post-training variants obtain 0% on OCR-related tasks, indicating that when the base model lacks reliable visual grounding, alignment alone cannot compensate for it. Instruction tuning may improve response style, but it does not restore missing perceptual capabilities.

> ***Finding 5.*** Standard SFT often weakens spatial grounding by emphasizing style over visual precision, while RL-based methods (RLHF/DPO) mitigate this effect to some extent, but the overall gains remain modest.

D.10 Ablation 10: Does scaling the language backbone in VLMs improve fundamental visual skills?

As shown in **Fig. 6**, models' results (Liu et al., 2023a; Li et al., 2024a; Bai et al., 2025a; Wang et al., 2025) on the FBI do not follow the standard scaling trend typically observed in language modeling. Increasing the parameter count of the language backbone does not lead to consistent improvements in multimodal spatial reasoning.

In fact, performance often stagnates or even declines as models scale up, suggesting an inverse scaling behavior in this setting. Larger language backbones appear to amplify linguistic priors without correspondingly strengthening visual grounding, which may exacerbate hallucination when tasks require precise spatial perception.

The performance collapse of large-scale models (e.g., 72B) demonstrates that blindly injecting tens of billions of parameters into the LLM backbone does not rectify fundamental weaknesses in geometric and spatial comprehension.

> ***Finding 6.*** Increasing the size of the language backbone does not improve fundamental capabilities; instead, larger models often rely more heavily on linguistic priors, leading to confident but spatially incorrect predictions.

D.11 Visualize

We add visualization of LLava in 7

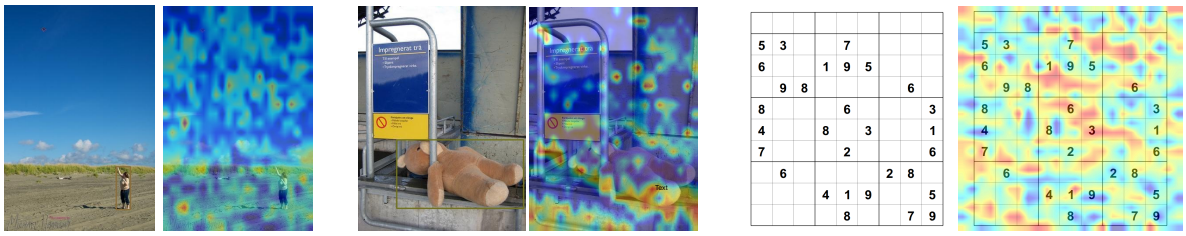
Table 19: Ablation 4: Evaluation results on spatial-centric datasets pretrained MLLMs regarding each capability integration. Accuracy is presented in %.

Model	Pretrained	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Llava-1.5-13B(Liu et al., 2023a)	None	3.6	2.9	0.0	4.8	4.0	29.2	4.4	0	0	12.5	0.0	0.0	0.0	4.2
Llava-1.5-13B(Liu et al., 2023a)	SpatialVLM(Chen et al., 2024a)	2.1	1.0	0.0	4.8	4.0	33.3	0.0	0	0	18.8	7.1	7.7	0.0	4.0
Llava-1.5-13B(Liu et al., 2023a)	COCO Caption(Chen et al., 2015)	8.6	8.8	0.0	2.4	12.0	33.3	4.4	0	0	0.0	0.0	7.7	0.0	7.0

Table 20: Evaluation results on LLaVA variants regarding each capability integration. Accuracy is presented in %.

Model	Loc Rel	Loc Rec Rel	Cou Rec	Dir Loc	Cou Loc OCR Rec	Com Loc	Rec	Cou Loc OCR	Cou OCR	Dir Rel	Dir Loc Rec	Dir Loc Rel	Loc Rec	Overall
Llava-1.5-7B(Liu et al., 2023a)	3.6	7.8	1.8	11.9	0.0	4.2	13.0	0.0	0.0	37.5	0.0	0.0	0.0	6.0
Llava-SFT-7B(von Werra et al., 2020)	2.1	1.0	0.0	0.0	0.0	20.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
Llava-RLHF-7B(Sun et al., 2023)	10.7	17.6	0.0	4.8	0.0	12.5	13.0	0.0	0.0	18.8	7.1	15.4	0.0	9.0
Llava-1.5-13B(Liu et al., 2023a)	3.6	2.9	0.0	4.8	4.0	29.2	4.3	0.0	0.0	12.5	0.0	0.0	0.0	4.0
Llava-1.5-13B-DPO(Li et al., 2024e)	8.6	9.8	0.0	2.4	4.0	16.7	4.3	0.0	0.0	18.8	7.1	0.0	0.0	7.0

Figure 6: FBI-Bench performance across different model scales.



Question: Where is the woman located?

Answer: The sand.

Models: The beach

(a)

Question: Which direction is the teddy bear facing?

Answer: Upward

Models: Left

(b)

Question: How many rows are there on this puzzle?

Answer: 10

Models: 9

(c)

Figure 7: Gradcam visualization of LLaVA-OneVision-7B on our benchmark.

E Data Annotation and Curation Protocol

This appendix provides the detailed annotation and curation protocol used to construct **FBI-Bench**. The goal is to make the benchmark construction process transparent, reproducible, and auditable. Since the main contribution of **FBI-Bench** is a carefully curated diagnostic benchmark for fundamental but currently unsolved multimodal skills, we document the data sources, annotator instructions, question-writing rules, filtering criteria, copyright considerations, and release format in detail.

E.1 Overview

FBI-Bench is designed to evaluate whether vision–language models can reliably solve human-easy questions that require fundamental multimodal perception. Each benchmark sample consists of an image, a free-form question, a canonical ground-truth answer, optional answer aliases, source metadata, and one or more capability labels. The benchmark focuses on seven foundational skills: recognition, localization, relation understanding, direction, counting, OCR, and comparison. A sample may require either a single skill or a composition of multiple skills.

The construction of **FBI-Bench** follows a three-stage curation pipeline. First, we collect images from public web sources and existing benchmarks, then filter them for visual clarity, source suitability, and relevance to fundamental perception. Second, annotators write free-form question–answer pairs, assign capability labels, and verify whether each candidate is sufficiently easy for humans. Third, we apply conservative filtering to remove duplicate, ambiguous, irrelevant, overly difficult, incorrectly labeled, or non-diagnostic samples, followed by model-based filtering to retain only examples that expose shared failures across current VLMs.

The pipeline follows four design principles.

- **Human-easy but model-hard.** Each retained question should be easy for humans to answer from the image, but difficult for current VLMs.
- **Perception-first.** The answer should be derivable primarily from visual evidence rather than external knowledge, memorization, or language priors.
- **Free-form evaluation.** The question should be answerable without seeing multiple-choice options, so that the model must generate the perceived answer rather than select from candidate choices.
- **Diagnostic annotation.** Each sample should be labeled with the required capability or capability composition, enabling fine-grained failure analysis.

E.2 Annotator Recruitment and Responsibilities

We recruited 10 annotators to construct and verify the benchmark. Annotators were instructed to collect diverse image candidates, write or refine questions, assign ground-truth answers, label required capabilities, and participate in multi-round quality-control voting. Before annotation, annotators received a written guideline describing the benchmark goal, allowed sources, copyright requirements, capability taxonomy, question-writing rules, filtering criteria, and examples of valid and invalid samples.

The annotators were assigned two functional roles.

- **Candidate collectors.** These annotators searched for visually diverse images and proposed initial question–answer pairs grounded in the image.
- **Verifiers.** These annotators independently checked whether each proposed question was

visually answerable, unambiguous, human-easy, correctly answered, and correctly labeled with the required capability or capability composition.

The same annotator was not allowed to be the only verifier of their own submitted sample. For every retained sample, at least one additional annotator checked the image, question, answer, source metadata, and capability labels. Disagreements were resolved by majority voting and, when necessary, discussion among annotators. Samples that remained controversial after discussion were removed.

E.3 Stage 1: Image Collection and Visual Filtering

The first stage focuses on collecting visually diverse candidate images and removing unsuitable images before question writing. At this stage, annotators do not yet focus on making the sample model-hard. Instead, the goal is to build a broad pool of visually clear, legally traceable, and perception-relevant images that can support human-solvable questions.

E.3.1 Image Sources

Candidate images were collected from two broad source groups: public web images and existing vision-language or visual reasoning benchmarks.

Public web images. Annotators collected images from publicly accessible web sources, including open web pages, educational websites, open-license image repositories, and search-engine results. Public web images are useful because they provide diverse real-world scenes, object layouts, text appearances, image styles, and visual conditions that may not be fully covered by existing benchmarks.

For every public web image, annotators were instructed to record the source URL, access date, source name when available, and license note when available. Images with explicit redistribution restrictions, paywalls, commercial watermarks, private-account origins, or unclear privacy status were excluded from the redistributable image set.

Images from existing benchmarks. We also considered images from existing vision-language and visual reasoning datasets, including COCO-style natural-image datasets, VCR-style visual reasoning datasets, and other public multimodal benchmarks. These images provide a controlled source of natural scenes and visual reasoning contexts.

When using images from prior benchmarks, annotators were instructed not to simply reuse the original question, caption, or answer. Instead, they inspected the image and wrote new free-form questions targeting the seven **FBI-Bench** foundational capabilities. For VCR-style images, annotators were specifically instructed to avoid questions requiring social commonsense, movie-specific context, character identity, or external knowledge. Only questions answerable from direct visual evidence were allowed.

E.3.2 Visual Filtering Criteria

After image collection, annotators performed visual filtering before writing questions. This step removes images that are visually unusable, legally problematic, or irrelevant to the benchmark goal.

An image is kept only if it satisfies the following criteria.

- **Visual clarity.** The image must be sufficiently clear and not severely blurred, corrupted, over-compressed, or too low-resolution.
- **Visible evidence.** The image must contain concrete visual evidence that can support at least one question about recognition, localization, relation, direction, counting, OCR, or comparison.
- **Perceptual relevance.** The image should support questions about basic perception rather than specialized domain knowledge, social inference, or external facts.
- **Source suitability.** The image must have traceable source metadata. Images with clear copyright, privacy, or redistribution concerns are removed or marked as non-redistributable.
- **Content diversity.** The image pool should cover diverse scenes, object categories, visual layouts, text appearances, and image styles to avoid source or content bias.

Images are removed at this stage if the target content is too small to inspect, the image is dominated by irrelevant visual clutter, the image requires domain-specific expertise to interpret, or the image lacks sufficient source information.

Source Type	Examples	Allowed Use	Required Metadata
Public web images	Open web pages, educational websites, open-license repositories, search-engine results	Allowed when the image is visually relevant, traceable, and legally usable	Source URL, access date, source name, license note
COCO-style images	Natural images containing people, animals, objects, signs, and common environments	Used for newly written free-form questions about fundamental perception	Dataset name, image ID, original split, license note if available
VCR-style images	Images originally used for visual commonsense reasoning	Used only when the new question can be answered from direct visual evidence	Dataset name, image ID, original split, license note if available
Other prior benchmarks	Existing VQA or visual reasoning datasets	Used only after rewriting into free-form diagnostic questions and relabeling capabilities	Dataset name, sample ID, original image ID, rewritten question

Table 21: Image source categories considered during **FBI-Bench** construction. Each candidate image is accompanied by source metadata to support reproducibility, auditing, and copyright review.

E.4 Stage 2: Question–Answer Annotation and Capability Labeling

The second stage converts visually valid images into candidate benchmark samples. For each image that passes Stage 1, annotators write one or more free-form question–answer pairs. The goal is to create questions that are simple for humans, directly grounded in the image, and aligned with the seven foundational capabilities.

E.4.1 Question-Writing Guidelines

Annotators were instructed to write questions that are simple, direct, and visually grounded. The question should test whether the model can perceive the relevant visual evidence, not whether it can infer an answer from answer choices, world knowledge, or language priors.

Valid questions must satisfy the following requirements.

- **The question must be answerable from the image alone.** The answer should not require external knowledge, hidden context, or dataset-specific information.
- **The question must be free-form.** Annotators should not rely on multiple-choice options, because answer choices may make the task easier through elimination or answer recognition.
- **The question must be short and direct.** The question should target a concrete visual property, object, relation, count, direction, text span, or comparison.
- **The answer must be short and checkable.** Valid answers are usually object names, colors, numbers, directions, positions, visible text spans, or short phrases.
- **The question must not leak the answer.** The wording should not contain hints that make the answer inferable without inspecting the image.
- **The sample must be human-easy.** A human annotator should be able to answer confidently from the image alone.

Annotators were instructed to reject or rewrite questions with the following problems.

- The answer requires external knowledge rather than visual evidence.
- The question is subjective, opinion-based, or dependent on personal interpretation.
- The image is too blurry, low-resolution, occluded, or cropped to support a reliable answer.
- Multiple answers are equally plausible.
- The question can be answered from language priors without inspecting the image.
- The question depends on recognizing a celebrity, private person, movie character, brand-specific trivia, or cultural context not visually explicit.
- The question is overly complex or requires long multi-step reasoning beyond fundamental perception.
- The wording leaks the answer.

Criterion	Valid Example	Invalid Example
Visual grounding	"What color is the umbrella held by the woman?"	"Why is the woman carrying an umbrella?"
Single answer	"How many dogs are on the sofa?"	"What is the most important object in the room?"
No external knowledge	"What word is written on the sign?"	"Which company owns the logo on the sign?"
Human-easy	"Is the cat on the left or right side of the box?"	"Estimate the exact distance between the cat and the box."
Capability-focused	"Which object is above the cup?"	"Describe everything happening in the scene."

Table 22: Examples of valid and invalid question-writing decisions. Valid questions are visually grounded, short, unambiguous, and aligned with fundamental perception.

E.4.2 Answer Normalization

Each question is paired with a canonical answer and, when needed, acceptable aliases. Answer normalization is used to reduce unnecessary variation during free-form evaluation.

- **Numeric answers.** Counts are stored using Arabic numerals as the canonical form, e.g., `3` rather than `three`.
- **OCR answers.** Visible text is transcribed as closely as possible. Minor case, whitespace, and punctuation variants may be included as aliases.
- **Direction answers.** Spatial and orientation answers use a normalized vocabulary such as `left`, `right`, `above`, `below`, `front`, `behind`, `upward`, and `downward`.
- **Object and attribute answers.** Common synonyms are included when they preserve the same meaning, such as `sofa` and `couch`.
- **Short phrase answers.** If the correct answer is a phrase, annotators record the shortest unambiguous phrase that preserves the intended meaning.

E.4.3 Capability Taxonomy

Each sample is labeled with one or more of the seven foundational capabilities.

- **Recognition (Rec).** Identify the presence, category, attribute, color, state, or type of a visual entity.
- **Localization (Loc).** Determine where an object, person, text region, or visual element is located in the image.
- **Relation (Rel).** Identify semantic or spatial relations between two or more entities.
- **Direction (Dir).** Determine orientation or ordered spatial predicates, such as left, right, above, below, front, behind, upward, or downward.
- **Counting (Cou).** Count entities satisfying a visual condition.
- **OCR (Ocr).** Read and transcribe text appearing in the image.
- **Comparison (Com).** Compare visual attributes, quantities, sizes, positions, lengths, or other magnitudes.

Annotators were instructed to label the minimal set of skills required to answer each question. The goal is not to label every possible skill visible in the image, but only the skills necessary for solving the question.

For example, if the question asks "What color is the car?", the correct label is `Rec`, even if the image also contains spatial relations. If the question asks "What color is the car to the left of the bus?", the correct labels are `Rec`, `Loc`, and `Rel`, because the annotator must identify the relevant car, locate the bus, determine the relation, and recognize the car color.

E.4.4 Human-Easy Verification

After writing each question–answer pair, annotators perform an initial human-easy check. A sample is kept only if the annotator believes that a human can answer the question confidently from the image alone. This step is important because **FBI-Bench** is intended to diagnose failures on

Label	Use When	Example Question
Rec	The answer requires recognizing an object, attribute, color, category, or state.	"What color is the girl's shirt?"
Loc	The answer requires locating an object or region in the image.	"Where is the dog located?"
Rel	The answer requires understanding the relation between two or more entities.	"What object is next to the cup?"
Dir	The answer requires orientation or ordered direction.	"Which direction is the arrow pointing?"
Cou	The answer requires counting entities satisfying a condition.	"How many red apples are on the table?"
Ocr	The answer requires reading visible text.	"What word is written on the sign?"
Com	The answer requires comparing size, number, position, length, or another visual property.	"Which box is larger?"

Table 23: Capability-labeling guideline used by annotators. Annotators assign the minimal set of capabilities needed to answer each question.

elementary perception, not failures caused by inherently difficult or ambiguous images.

A sample fails the human-easy check if:

- the target object is too small, blurry, or occluded for reliable human recognition;
- the image does not contain enough evidence to determine the answer;
- the question requires specialized knowledge or hidden context;
- the expected answer depends on subjective interpretation;
- multiple answers are plausible;
- the question requires complex reasoning beyond fundamental perception.

Samples failing this check are either rewritten or discarded before entering the final filtering stage.

E.5 Stage 3: Multi-Round Filtering and Final Selection

The third stage removes invalid, ambiguous, duplicate, irrelevant, overly difficult, or non-diagnostic samples. This stage is intentionally conservative because the final benchmark is meant to expose model failures on human-easy samples rather than artifacts caused by noisy annotation.

E.5.1 Filtering Criteria

We apply the following filtering steps.

Duplicate filtering. We remove repeated images, near-duplicate images, repeated questions, and samples that test the same visual evidence with nearly identical wording. Near-duplicates are identified using filename matching, source metadata, and manual visual inspection.

Ambiguity filtering. We remove or rewrite samples where multiple answers are plausible, the target object is unclear, the referring expression is underspecified, or annotators disagree about the answer. Ambiguity is treated conservatively: if a sample remains controversial after discussion, it is removed.

Irrelevance filtering. We remove samples that do not directly test the seven target capabilities, require external knowledge, depend on subjective interpretation, or rely mainly on language priors rather than image evidence.

Difficulty filtering. We remove samples that are too difficult for humans because of extreme blur, tiny targets, unreadable text, severe occlusion, insufficient resolution, or overly complex reasoning requirements. This step ensures that model failure cannot be attributed simply to impossible or low-quality images.

Ground-truth verification. We check whether the canonical answer and aliases are correct, complete, and consistent with the image. Samples with incorrect or incomplete ground truth are corrected or removed.

Capability-label verification. We check whether the assigned capability labels represent the mini-

mal set of skills required to answer the question. Samples with unstable or disputed capability labels are relabeled or discarded.

Source and license verification. We check whether the sample has sufficient source metadata and whether the image can be redistributed or should be released only through metadata or source references.

E.5.2 Model-Based Filtering

After human filtering, we apply model-based filtering to identify samples that reveal shared limitations across current VLMs. Candidate samples are evaluated with multiple VLMs under the same free-form setting used in the main evaluation. If at least one filtering model answers a sample correctly, that sample is excluded from the final FBI-Bench set. Only samples failed by all filtering models are retained.

This filtering step is central to the design of FBI-Bench. A sample failed by only one model may reflect model-specific weakness, stochastic error, or implementation artifacts. In contrast, a sample failed consistently across diverse models is more likely to reveal a shared limitation in current VLMs. Therefore, model-based filtering helps ensure that the final benchmark emphasizes systematic failures rather than isolated model-specific mistakes.

E.5.3 Final Manual Sanity Check

After model-based filtering, annotators conduct a final manual sanity check over the retained samples. This final pass verifies:

- the image is visually clear enough for humans;
- the question is answerable from the image alone;
- the answer and aliases are correct;
- the capability labels are accurate;
- the source metadata is complete when available;
- the sample does not contain private or sensitive information;
- the formatting is consistent with the release schema.

Only samples passing this final check are included in the released benchmark.

Stage	Curation Process	Samples	Retention
Initial	Raw image candidates and image-question candidates (X_0)	24,460	100%
Stage 1	Image collection and visual filtering (X_1)	18,285	74.8%
Stage 2	QA annotation, capability labeling, and human-easy verification (X_2)	10,142	41.5%
Stage 3a	Duplicate, ambiguity, irrelevance, and difficulty filtering (X_3)	5,318	21.7%
Stage 3b	**Model-based filtering and final sampling**	**500**	**9.4% of X_3**

Table 24: Three-stage curation funnel of FBI-Bench. Stage 1 filters visually suitable images, Stage 2 constructs and verifies question–answer annotations, and Stage 3 removes noisy or non-diagnostic samples before model-based filtering and final selection.

E.6 Ambiguity Resolution

Ambiguity was treated conservatively throughout annotation. If a question could reasonably have more than one answer, the sample was removed or rewritten. Annotators were instructed to flag the following cases:

- the target object is too small, blurry, partially hidden, or visually indistinguishable;
- multiple objects satisfy the same referring expression;
- the question uses vague terms such as “near”, “large”, “small”, or “main” without a clear visual criterion;
- the answer depends on subjective judgment;
- the image contains text that is partially readable but not fully reliable;
- different annotators select different ground-truth answers.

When ambiguity could be resolved by rewriting the question, annotators rewrote it to include a clearer reference. For example, “What is the object near the man?” could be rewritten as “What object is directly to the left of the man in the blue shirt?” If ambiguity remained after rewriting, the sample was discarded.

E.7 Copyright, Privacy, and Source-Use Policy

Since FBI-Bench is curated from public internet sources and existing datasets, we apply a conservative source-use policy.

- **No private or sensitive images.** Annotators were instructed not to collect images from private accounts, personal social media pages, medical records, surveillance footage, or sources containing personally sensitive information.

- **No clearly restricted images.** Images with explicit no-redistribution statements, commercial watermarks, paywall restrictions, or unclear ownership were excluded from the redistributable image set.

- **License tracking.** For every candidate image, annotators recorded source information whenever available, including the source URL, source dataset, image ID, and license note.

- **Dataset-derived images.** For images from COCO, VCR, or other datasets, we preserve original dataset identifiers so that users can trace the sample back to the original source.

- **Release separation.** When redistribution rights are clear, we release the image together with annotation metadata. When redistribution rights are uncertain, we release only annotation metadata, source URL, dataset identifier, and reconstruction instructions where permitted.

This policy is intended to make the benchmark useful for the community while respecting copyright, licensing, and privacy constraints.

E.8 Annotation Format and JSON Schema

Each accepted sample is stored as a JSON record. The schema is designed to support evaluation, source auditing, capability analysis, and future community inspection.

Sample JSON Schema

```
{
  "sample_id": "fbi_000001",
  "image_path": "images/fbi_000001.jpg",
  "source_type": "COCO",
  "source_name": "...",
  "source_url": "...",
  "source_image_id": "...",
  "license_note": "...",
  "question": "...",
  "answer": "...",
  "answer_aliases": ["...", "..."],
  "capabilities": ["Loc", "Rel"],
  "capability_integration": "Loc-Rel",
  "difficulty_note": "human-easy",
  "ambiguity_flag": false,
  "annotator_id": "ann_03",
  "verifier_ids": ["ann_05", "ann_08"],
  "verification_result": "accepted",
  "model_filtering_result": {
    "all_filtering_models_failed": true,
    "models_solved": []
  }
}
```

The field `source_type` takes one of four values: `public_web`, `COCO`, `VCR`, or `prior_benchmark`. The field `capabilities` stores the atomic skill labels, while `capability_integration` stores the composed label used for analysis. The field `model_filtering_result` records whether the sample survived model-based rejection.

E.9 Human Validation

To confirm that the benchmark is human-solvable, we conduct human validation after the main filtering stages. Participants from multiple education levels answer the questions under the same free-form setting used for model evaluation. The purpose of this validation is not to make the benchmark difficult for humans, but to verify that the samples are well-posed and that model failure is not caused by annotation noise.

Samples with unexpectedly low human agreement are manually reviewed. If the error is caused by ambiguity, poor image quality, unclear wording, or incorrect ground truth, the sample is removed or corrected. This validation step supports the central design goal of **FBI-Bench**: the final samples should be easy for humans but systematically difficult for current VLMs.

E.10 Release Format and Community Use

We plan to release the benchmark annotations, evaluation code, and source metadata to support reproducibility and community use. The released files will include:

- the image or image reference, depending on redistribution permission;

- the free-form question;

- the canonical answer and acceptable aliases;

- the capability labels and capability-integration label;

- source metadata, including dataset name, source URL, or original image ID when available;

- license notes or redistribution notes;

- evaluation scripts for open-ended model outputs;

- the LLM-judge prompt and answer-normalization rules.

For images that cannot be redistributed directly, we will release metadata and source references instead of the raw image file, when permitted. This allows the community to inspect and reconstruct the benchmark while respecting source restrictions. We also plan to provide a mechanism for users to report questionable samples, including possible ambiguity, incorrect labels, licensing concerns, or broken source links.

E.11 Summary

This protocol clarifies how **FBI-Bench** is constructed from public web images and prior benchmark sources. The annotation process follows three stages: image collection and visual filtering, question–answer annotation with capability labeling, and multi-round filtering with final model-based selection. Together, these steps ensure that the final benchmark is not merely a collection of difficult examples, but a compact diagnostic set for revealing systematic failures in the fundamental perceptual abilities of current VLMs.

F Failure Taxonomy: Annotation Protocol and Category Definitions

F.1 Overview

The failure taxonomy used in **Sec. 4.3** was constructed through a structured human annotation process designed to identify the primary reason a sample is difficult for a model to answer correctly, independent of which specific model is being evaluated. The goal is not to explain why a particular architecture fails on a particular instance, but to characterise the nature of the visual or reasoning demand that systematically defeats current models as a class. This distinction matters because a sample can be hard for idiosyncratic reasons, such as an unusual prompt phrasing or a model-specific representation gap, and such samples would not reveal anything informative about the field-level bottleneck. The taxonomy is therefore designed to surface only those failure modes that are structural and shared.

Each sample in both FBI and FBI-Lite was independently annotated by two annotators drawn from the same pool used during benchmark construction. Where the two annotators disagreed, a third annotator adjudicated. The final label is the majority decision across the three annotations. Inter-annotator agreement on the primary category reached 81.4% before adjudication, indicating that the taxonomy is well-defined enough to be applied consistently across annotators without extensive coordination.

Annotators were instructed to assign exactly one primary category per sample, selecting the failure mode that most directly explains why the question is structurally hard for a model to answer correctly, not why any specific model happened to fail on it. Alongside the category label, annotators recorded a free-text justification that was used during a calibration round to audit consistency and to sharpen the boundaries between adjacent categories, particularly between spatial imagination and prior-knowledge overwriting, which share the property that relevant visual evidence is present in the image but is not correctly used.

F.2 Annotation Protocol

The annotation followed a fixed five-step procedure applied uniformly across all samples.

In the first step, annotators read the question and the ground-truth answer carefully to understand what visual evidence in the image is necessary and sufficient to arrive at the correct response. This step ensures that the subsequent image inspection is directed rather than exploratory.

In the second step, annotators examined the image independently and attempted to answer the question before reading any model output. This deliberate sequencing anchors the annotation in the visual content of the image and prevents model outputs from biasing the annotator's judgment about what the image actually shows.

In the third step, annotators identified the single most challenging demand the question places on a system that processes images through a patch-based encoder followed by a language decoder. This framing focuses attention on architectural difficulty rather than on incidental image properties, and it encourages annotators to reason about the class of model being evaluated rather than about any specific failure instance.

In the fourth step, annotators assigned one of the five categories described below. When a sample plausibly belonged to two categories, annotators were instructed to select the one that, in their judgment, is more directly responsible for the failure at the level of mechanism rather than at the level of surface symptom. This rule most frequently applies at the boundary between spatial imagination and prior-knowledge overwriting, and the calibration round produced concrete guidance for disambiguat-

ing these cases.

In the fifth step, annotators wrote a one-to-two sentence justification. These justifications were reviewed collectively during calibration and used to produce a revised annotation guide that was distributed before the main annotation pass.

Before the main pass, all annotators completed a calibration round on 50 shared samples. Initial pairwise agreement on those samples was 71%, and collective discussion of the disagreements raised it to 81.4% on the subsequent main pass. The calibration round also produced a set of boundary cases that serve as reference points for the most frequently confused category pairs.

F.3 Category Definitions

F.3.1 Low-Level Visual Confounders

This category captures cases where the primary obstacle is a physical property of the image that impedes access to the relevant visual evidence, regardless of the reasoning required once that evidence is obtained. The conceptual demand of the question is straightforward; the difficulty lies in extracting reliable features from the image in the first place.

We distinguish three sub-types. Small objects occur when the target entity occupies a small fraction of the image and its visual features are consequently under-represented in the patch-level feature map. Standard patch sizes aggregate small targets with surrounding background, effectively erasing their discriminative features before they reach the language module. This is a well-known limitation of fixed-resolution vision encoders and is expected to affect performance independently of the reasoning capability of the language backbone.

Occlusion occurs when a significant portion of the target entity is hidden behind another object in the scene. The visible portion may be insufficient to determine the entity's identity, colour, orientation, or count with confidence. Partial occlusion of this kind does not necessarily make a sample unanswerable for a human who can infer the hidden portions from context, but it substantially reduces the quality of the visual signal available to a model that relies on dense feature extraction.

High density occurs when the scene contains many similar objects in close proximity, creating a situation where object boundaries are ambiguous and individual instances are difficult to separate. This sub-type most commonly appears in counting tasks, where the difficulty of individuating objects in a dense arrangement makes precise enumeration unreliable even for models with strong spatial representations.

A sample is assigned to this category only when image-level difficulty is genuinely the bottleneck and the reasoning demand placed on the model after visual evidence is obtained is minimal. If a dense scene requires a relational inference between two specific objects once they are localised, the dominant failure mode is spatial imagination rather than scene density.

F.3.2 Spatial Imagination

This category captures failures that arise when answering correctly requires the model to perform a spatial transformation or inference that goes beyond reading feature values directly from the image. Unlike low-level confounders, the relevant visual evidence is present and accessible in the image; the difficulty lies in what must be done with that evidence to produce the correct answer.

We distinguish three sub-types. Viewpoint shift occurs when the question asks about a spatial relationship from a reference frame that differs from the camera viewpoint. Questions of the form "if you were standing at position X facing direction Y, where would object Z be relative to you" require the model to mentally reorient the scene, inverting or rotating spatial relationships that are visible from the camera but not from the queried perspective. This transformation requires maintaining a representation of the scene's three-dimensional structure that is decoupled from the two-dimensional image plane, a capability that current patch-based encoders do not appear to support reliably.

Pose and orientation failures occur when the question asks about the facing direction, body orientation, or posture of an entity whose pose must be inferred from the geometry of visible body parts or object components. Determining whether a person is facing toward or away from the camera, which direction a vehicle's front is pointing, or whether an object's labeled side faces left or right in a given image requires interpreting the spatial arrangement of parts rather than reading a single visual feature. Models that process images as bags of local features rather than structured spatial representations are expected to struggle with this sub-type systematically.

Spatial relation inference occurs when the question asks about the relative position of two or more entities in a way that requires integrating multiple

localisation decisions and reasoning about their geometric relationship. This sub-type is among the most frequent in FBI and is responsible for a large share of the gap between FBI-Lite and FBI. Unlike simple localization, which asks where a single object is, spatial relation inference requires the model to simultaneously maintain representations of multiple entities and to compute a relational predicate over them, a compositional demand that current architectures handle inconsistently.

The boundary between spatial imagination and prior-knowledge overwriting is the most frequently contested in annotation. The distinguishing criterion is whether the correct answer requires spatial reasoning that is genuinely non-trivial, meaning that a careful human observer must perform a mental transformation to arrive at it. If the spatial relationship is directly readable from the image without transformation and the model nonetheless answers incorrectly by relying on a prior, the sample belongs to prior-knowledge overwriting.

F.3.3 Prior-Knowledge Overwriting

This category captures cases where the model has sufficient visual evidence to answer correctly but substitutes a response generated from a linguistic prior or a scene-level heuristic. The failure is not perceptual in the sense of missing visual information; it is a grounding failure, where the language module overrides the visual signal with a statistically likely response that the training distribution makes salient.

We distinguish two sub-types. Superficial grounding occurs when the model identifies the scene category or the dominant object class and generates an answer based on that category-level identification rather than by localising the specific entity or attribute the question queries. A model exhibiting superficial grounding is answering a different, easier question than the one posed: instead of determining where within the image a specific person is standing, it reports what type of scene the image depicts. This behaviour is consistent with a model that has learned strong associations between scene categories and likely answers during instruction tuning, and that activates those associations before completing fine-grained visual analysis.

Evidential perception failures occur when the model attends to a salient but incorrect visual cue and ignores a less salient cue that is diagnostic for the correct answer. This sub-type differs from superficial grounding in that the model does engage with the image at some level of detail but selectively weights features in a way that is misaligned with the question's requirements. The dominant salient feature is not the scene category but a specific object or attribute that is visually prominent and primes an incorrect response.

The practical consequence of this category for the FBI versus FBI-Lite comparison is significant. The rise in superficial grounding from 2.2% in FBI-Lite to 22.2% in FBI reveals that a substantial fraction of the samples that at least one model solves in the pre-filtered pool are solved not through genuine visual understanding but through scene-level shortcuts. Once the benchmark is filtered to remove these samples, the shortcut-reliant success rate disappears entirely, and what remains is a set of questions that genuinely require the model to ground its answer in fine-grained visual evidence.

F.3.4 Soft Cognitive Variance

This category captures errors that arise from genuine perceptual ambiguity in the image rather than from a structural model limitation. Samples in this category are ones where a careful human observer might also make errors, though typically at a much lower rate than current models, and where the failure therefore does not constitute reliable evidence of a fundamental architectural deficit.

We distinguish three sub-types. Counting variance occurs when the exact number of a set of objects is difficult to determine because objects partially overlap, are arranged in complex patterns, or are numerous enough that small errors of enumeration are plausible even under careful inspection. The difficulty here is not in the concept of counting but in the precision of individuating objects in a scene where boundaries are not cleanly separated.

Attribute confusion occurs when an object's property, such as its colour, texture, or material, is genuinely ambiguous due to lighting conditions, shadow, reflection, or image compression artefacts. The correct answer exists and is determinable by careful inspection, but the visual signal is noisy enough that errors are understandable rather than indicative of a systematic failure mode.

Text misreading occurs when the text visible in the image is degraded, stylised, partially occluded, or rendered in an unusual font, making accurate OCR-dependent answers non-trivial even for a competent reader. Unlike the OCR failures that fall under prior-knowledge overwriting, where the text is legible but the model ignores it in favour

of a prior, here the difficulty originates in the image itself.

The most important observation about this category in the context of the FBI versus FBI-Lite comparison is its complete disappearance in FBI. In FBI-Lite, soft cognitive variance accounts for 36.6% of annotated failure modes, suggesting that a large share of the samples that models occasionally solve are ones where the correct answer can be arrived at through imprecise or lucky attribute guessing. Once these samples are filtered out, the remaining set in FBI contains no instances of this type, confirming that the filtered benchmark tests for precise, structurally grounded understanding rather than for approximate perceptual judgments.

F.3.5 Others

This category is reserved for samples that do not fit cleanly into any of the four categories above. In practice it comprises two sub-populations. The first consists of samples where multiple failure modes contribute roughly equally and no single one is primary; annotators were instructed to avoid assigning this label when one mode was clearly dominant, so its presence indicates genuine multi-causal cases. The second sub-population consists of rare cases where domain-specific knowledge is required to understand what the question is asking, placing the failure outside the scope of perceptual or spatial capability and into the domain of factual grounding.

The frequency of this category declines sharply from 16.8% in FBI-Lite to 1.6% in FBI. This decline is consistent with the filtering operation having removed samples whose difficulty does not arise from the structured failure modes the taxonomy is designed to capture. The near-zero rate in FBI confirms that the filtered benchmark is dominated by the four primary categories and that multi-causal or domain-knowledge cases are not a confound in the failure analysis.

F.4 Taxonomy Validation

To validate that the taxonomy separates structurally distinct failure types rather than producing arbitrary groupings, we conducted two post-hoc checks.

For the first check, we tested whether the category distributions differ significantly between FBI and FBI-Lite using a chi-squared test across all five categories. The difference is highly significant ($\chi^2 = 312.4$, $p < 0.001$), confirming that the two sets are not drawn from the same failure distribution and that the taxonomy is sensitive to the filtering operation that separates them. This result also validates the use of FBI-Lite as a meaningful contrast set: it is not merely a random subset of the same population but a set with a structurally distinct failure profile.

For the second check, we asked a separate group of three annotators who had not participated in the original annotation to re-label a random sample of 100 items from FBI using only the category definitions above, without access to the original labels. Agreement with the original labels reached 79.3%, comparable to the inter-annotator agreement during the main pass. This result confirms that the taxonomy can be applied consistently by new annotators using the written definitions alone, without the benefit of the calibration round that the original annotators completed. It also provides evidence that the taxonomy is not over-fitted to the judgments of the original annotator pool and generalises to independent raters.

G Qualitative Analysis

List of Failure Cases

1. **Counting · Location · OCR** *How many circles are there in cell E5?*
2. **Counting · OCR** *If set A includes elements that intersect with set B, how many elements are there in total?*
3. **Counting · OCR** *How many data points are above 40?*
4. **Counting · Recognition** *How many yellow bears are there?*
5. **Counting · Recognition** *How many apples and bananas are in the image?*
6. **Direction · Location** *Which object is the bicycle facing? (1)*
7. **Direction · Location** *Which object is the bicycle facing? (2)*
8. **Direction · Location** *In which direction is the teddy bear facing?*
9. **Direction · Relation** *Which object is the small horse facing?*
10. **Location · Recognition · Relation** *Are the orientations of the giraffes parallel or perpendicular?*
11. **Location · Recognition · Relation** *Which side of the girl in the red dress is facing the beer bottle?*
12. **Location · Recognition · Relation** *From the TV's perspective, which side of the man in white is facing it?*
13. **Location · Recognition · Relation** *Is the man in black directly underneath the blue umbrella?*
14. **Location · Recognition · Relation** *Is the old man directly underneath the umbrella?*
15. **Location · Relation** *From the perspective of someone standing at the shrine, which side of the couch?*
16. **Location · Relation** *Where is the woman located?*
17. **Location · Relation** *Which side of the person is facing the chair?*
18. **Location · Relation** *Where is the man positioned in relation to the woman?*
19. **Location · Relation** *Where is the woman located in relation to the man?*
20. **Location · Relation** *From the perspective of the window, which side of the lady is visible?*
21. **Location · Relation** *Where is the teapot located?*
22. **Location · Relation** *Describe the boy's position relative to the sheep.*
23. **Location · Relation** *Where is the oven located in relation to the sink?*
24. **Recognition** *Considering their size, how do the pizza cutter and scissors compare?*
25. **Recognition** *Is the net in the background blue, red, or green?*
26. **Recognition** *What are the two boys wearing white clothes doing instead of skating?*

Capabilities: Counting · Location · OCR

Open-ended Question

Question: How many circles are there in cell E5? Answer with a number in curly brackets, e.g., 9.

Answer: 2

GPT-4o: 3
GPT-4o-mini: 5
Claude-haiku-4.5: 5
Claude-sonnet-4.5: 3
360VL-70B: 9
InternVL2_5-78B: 5
InternVL3-78B: 9
Kimi-A3B-Instruct: 8
Qwen2.5-72B-Instruct: 5
Qwen3-30B-A3B-Instruct: 1
Llava-OneVision-72B: 1

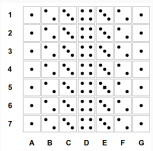
Multiple Choice Question

Question: How many circles are there in cell E5? Answer with a number in curly brackets, e.g., 9. (A) 2 (B) 1 (C) 3 (D) 4

Answer: A

Intern3.5-8B: C
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

Image Input



↑ Back to List of Failure Cases

Capabilities: Counting · OCR

Open-ended Question

Question: If set A includes elements that intersect with set B, how many elements are there in total?

Answer: 60

GPT-4o: 200
GPT-4o-mini: 200
Claude-haiku-4.5: 260
Claude-sonnet-4.5: 200
360VL-70B: 250
InternVL2_5-78B: 200
InternVL3-78B: 200
Kimi-A3B-Instruct: 140
Qwen2.5-72B-Instruct: 260
Qwen3-30B-A3B-Instruct: 320
Llava-OneVision-72B: 140

Multiple Choice Question

Question: If set A includes elements that intersect with set B, how many elements are there in total? (A) 60 (B) 80 (C) 140 (D) 200

Answer: A

Intern3.5-8B: A
Kimi-A3B: C
Qwen3-8B: C
Llava-OneVision-7B: C
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Counting · OCR

Open-ended Question

Question: How many data points are above 40?

Answer: 16

GPT-4o: 10
GPT-4o-mini: 13
Claude-haiku-4.5: 8
Claude-sonnet-4.5: 6
360VL-70B: 0
InternVL2_5-78B: 3
InternVL3-78B: 4
Kimi-A3B-Instruct: 2
Qwen2.5-72B-Instruct: 10
Qwen3-30B-A3B-Instruct: 2
Llava-OneVision-72B: 6

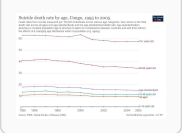
Multiple Choice Question

Question: How many data points are above 40? (A) 16 (B) 14 (C) 19 (D) 23

Answer: A

Intern3.5-8B: C
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: D
Internvl3-78B: A

Image Input



↑ Back to List of Failure Cases

Capabilities: Counting · Recognition

Open-ended Question

Question: How many yellow bears are there?

Answer: 5

GPT-4o: Four.
GPT-4o-mini: 6
Claude-haiku-4.5: 3
Claude-sonnet-4.5: 4
360VL-70B: 0
InternVL2_5-78B: 2
InternVL3-78B: 3
Kimi-A3B-Instruct: 0
Qwen2.5-72B-Instruct: 0
Qwen3-30B-A3B-Instruct: 2
Llava-OneVision-72B: 0

Multiple Choice Question

Question: How many yellow bears are there? (A) 3 (B) 5 (C) 4 (D) 6

Answer: B

Intern3.5-8B: D
Kimi-A3B: D
Qwen3-8B: D
Llava-OneVision-7B: D
Internvl3-78B: D

Image Input

↑ Back to List of Failure Cases

Capabilities: Counting · Recognition

Open-ended Question

Question: How many apples and bananas are in the image?

Answer: 1 apples and 5 bananas

GPT-4o: 4 bananas, 2 apples.
GPT-4o-mini: 2 apples and 7 bananas
Claude-haiku-4.5: 0 apples and 6 bananas
Claude-sonnet-4.5: 0 apples and 7 bananas
360VL-70B: 4 bananas 1 apple
InternVL2_5-78B: 6
InternVL3-78B: There is 1 apple and 6 bananas in the image.
Kimi-A3B-Instruct: 0
Qwen2.5-72B-Instruct: 2 apples, 4 bananas
Qwen3-30B-A3B-Instruct: There are 2 apples and 6 bananas in the image.
Llava-OneVision-72B: 10

Multiple Choice Question

Question: How many apples and bananas are in the image? (A) 2 apples and 5 bananas (B) 1 apples and 5 bananas (C) 3 apples and 5 bananas (D) 4 apples and 5 bananas

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: B
Llava-OneVision-7B: B
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Direction · Location

Open-ended Question

Question: Which object is the bicycle facing?

Answer: man

GPT-4o: A postbox.
GPT-4o-mini: postbox
Claude-haiku-4.5: The bicycle is facing the yellow/orange building.
Claude-sonnet-4.5: The bicycle is facing the Fullingers building.
360VL-70B: Mailbox
InternVL2_5-78B: Post box
InternVL3-78B: The post.
Kimi-A3B-Instruct: post office box.
Qwen2.5-72B-Instruct: post box
Qwen3-30B-A3B-Instruct: The bicycle is facing the car.
Llava-OneVision-72B: Mailbox

Multiple Choice Question

Question: Which object is the bicycle facing? (A) man (B) postbox (C) car (D) building

Answer: A

Intern3.5-8B: C
Kimi-A3B: D
Qwen3-8B: D
Llava-OneVision-7B: D
Internvl3-78B: B

Image Input

↑ Back to List of Failure Cases

Capabilities: Direction · Location

Open-ended Question

Question: Which object is the bicycle facing?

Answer: man

GPT-4o: Post box.
GPT-4o-mini: post
Claude-haiku-4.5: Fire hydrant
Claude-sonnet-4.5: The orange traffic cone.
360VL-70B: Mailbox
InternVL2_5-78B: Mailbox
InternVL3-78B: The post.
Kimi-A3B-Instruct: post office box.
Qwen2.5-72B-Instruct: post box
Qwen3-30B-A3B-Instruct: The bicycle is facing the road.
Llava-OneVision-72B: Mailbox

Multiple Choice Question

Question: Which object is the bicycle facing? (A) man (B) trash can (C) car (D) building

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Direction · Location

Open-ended Question

Question: In the picture, in which direction is the teddy bear facing?

Answer: upward

GPT-4o: Left
GPT-4o-mini: left
Claude-haiku-4.5: The teddy bear is facing downward/toward the ground.
Claude-sonnet-4.5: The teddy bear is facing to the right.
360VL-70B: Left
InternVL2_5-78B: Left
InternVL3-78B: Left.
Kimi-A3B-Instruct: left
Qwen2.5-72B-Instruct: left
Qwen3-30B-A3B-Instruct: The teddy bear is facing upwards.
Llava-OneVision-72B: Left

Multiple Choice Question

Question: In the picture, in which direction is the teddy bear facing? (A) upward (B) downward (C) to the right (D) toward the sign

Answer: A

Intern3.5-8B: B
Kimi-A3B: C
Qwen3-8B: C
Llava-OneVision-7B: C
Internvl3-78B: D

Image Input

↑ Back to List of Failure Cases

Capabilities: Direction · Relation

Open-ended Question

Question: Which object is the small horse facing?

Answer: tree

GPT-4o: The small horse is facing the car.
GPT-4o-mini: man
Claude-haiku-4.5: The small horse is facing the car.
Claude-sonnet-4.5: The car.
360VL-70B: Car
InternVL2_5-78B: The small horse is facing the person walking it.
InternVL3-78B: The small horse is facing the road.
Kimi-A3B-Instruct: dog.
Qwen2.5-72B-Instruct: person
Qwen3-30B-A3B-Instruct: The small horse is facing the dog.
Llava-OneVision-72B: Car

Multiple Choice Question

Question: Which object is the small horse facing? (A) tree (B) car (C) man (D) fence

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Recognition · Relation

Open-ended Question

Question: Are the orientations of the giraffes, one in front and one in back, parallel or perpendicular to each other?

Answer: perpendicular

GPT-4o: Parallel
GPT-4o-mini: parallel
Claude-haiku-4.5: Parallel
Claude-sonnet-4.5: Parallel
360VL-70B: Parallel
InternVL2_5-78B: Parallel
InternVL3-78B: Parallel
Kimi-A3B-Instruct: Parallel.
Qwen2.5-72B-Instruct: Parallel
Qwen3-30B-A3B-Instruct: parallel
Llava-OneVision-72B: Parallel

Multiple Choice Question

Question: Are the orientations of the giraffes, one in front and one in back, parallel or perpendicular to each other? (A) perpendicular (B) parallel (C) facing the same direction (D) facing opposite directions

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: B

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Recognition · Relation

Open-ended Question

Question: Which side of the girl in the red dress is facing the beer bottle?

Answer: front

GPT-4o: Her right side.
GPT-4o-mini: Right
Claude-haiku-4.5: Left side
Claude-sonnet-4.5: Left
360VL-70B: Right
InternVL2_5-78B: Left
InternVL3-78B: Left.
Kimi-A3B-Instruct: left
Qwen2.5-72B-Instruct: left
Qwen3-30B-A3B-Instruct: The girl's left side is facing the beer bottle.
Llava-OneVision-72B: Left

Multiple Choice Question

Question: Which side of the girl in the red dress is facing the beer bottle? (A) front (B) left (C) right (D) back

Answer: A

Intern3.5-8B: C
Kimi-A3B: C
Qwen3-8B: C
Llava-OneVision-7B: C
Internvl3-78B: C

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Recognition · Relation

Open-ended Question

Question: From the perspective of the TV, which side of the man in the white t-shirt is facing it?

Answer: back

GPT-4o: Left side.
GPT-4o-mini: left
Claude-haiku-4.5: Left side
Claude-sonnet-4.5: Left side
360VL-70B: Left
InternVL2_5-78B: Left
InternVL3-78B: Left
Kimi-A3B-Instruct: left
Qwen2.5-72B-Instruct: left
Qwen3-30B-A3B-Instruct: Left
Llava-OneVision-72B: Left

Multiple Choice Question

Question: From the perspective of the TV, which side of the man in the white t-shirt is facing it? (A) back (B) front (C) left (D) right

Answer: A

Intern3.5-8B: C
Kimi-A3B: C
Qwen3-8B: A
Llava-OneVision-7B: C
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Recognition · Relation

Open-ended Question

Question: Is the man in black directly underneath the blue umbrella?

Answer: yes

GPT-4o: No.
GPT-4o-mini: No
Claude-haiku-4.5: No.
Claude-sonnet-4.5: No.
360VL-70B: No
InternVL2_5-78B: No
InternVL3-78B: No.
Kimi-A3B-Instruct: No.
Qwen2.5-72B-Instruct: No
Qwen3-30B-A3B-Instruct: No
Llava-OneVision-72B: No

Multiple Choice Question

Question: Is the man in black directly underneath the blue umbrella? (A) yes (B) no (C) no — he's sitting behind the table (D) no — he's off to the side, not under the umbrella

Answer: A

Intern3.5-8B: B
Kimi-A3B: B
Qwen3-8B: B
Llava-OneVision-7B: B
Internvl3-78B: D

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Recognition · Relation

Open-ended Question

Question: Is the old man directly underneath the umbrella?

Answer: yes

GPT-4o: No.
GPT-4o-mini: No
Claude-haiku-4.5: No.
Claude-sonnet-4.5: No
360VL-70B: No
InternVL2_5-78B: No
InternVL3-78B: No.
Kimi-A3B-Instruct: No.
Qwen2.5-72B-Instruct: No
Qwen3-30B-A3B-Instruct: No
Llava-OneVision-72B: No

Multiple Choice Question

Question: Is the old man directly underneath the umbrella? (A) yes (B) no (C) he is standing next to the umbrella (D) no — he is in direct sunlight

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: From the perspective of someone standing at the shrine, which side of the couch would they be looking at?

Answer: front

GPT-4o: Right side.
GPT-4o-mini: back
Claude-haiku-4.5: The back of the couch.
Claude-sonnet-4.5: The back of the couch.
360VL-70B: Left
InternVL2_5-78B: The back.
InternVL3-78B: The back.
Kimi-A3B-Instruct: right
Qwen2.5-72B-Instruct: back
Qwen3-30B-A3B-Instruct: back
Llava-OneVision-72B: Back

Multiple Choice Question

Question: From the perspective of someone standing at the shrine, which side of the couch would they be looking at? (A) front (B) back (C) left side (D) right side

Answer: A

Intern3.5-8B: B
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: D

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: Where is the woman located?

Answer: The woman is on the bottom right

GPT-4o: Beach
GPT-4o-mini: On the beach
Claude-haiku-4.5: The woman is located on a beach.
Claude-sonnet-4.5: The woman is located on a beach.
360VL-70B: Beach
InternVL2_5-78B: Beach
InternVL3-78B: On a beach.
Kimi-A3B-Instruct: Beach.
Qwen2.5-72B-Instruct: beach
Qwen3-30B-A3B-Instruct: The woman is located on a sandy beach.
Llava-OneVision-72B: Beach

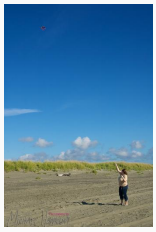
Multiple Choice Question

Question: Where is the woman located? (A) The woman is on the bottom right (B) The woman is on the top left (C) The woman is in the center (D) The woman is on the bottom left

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: C

Image Input



↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: Which side of the person is facing the chair?

Answer: back

GPT-4o: Left side.
GPT-4o-mini: Right
Claude-haiku-4.5: Right side.
Claude-sonnet-4.5: The right side of the person is facing the chair.
360VL-70B: Right
InternVL2_5-78B: Left
InternVL3-78B: Left.
Kimi-A3B-Instruct: Right.
Qwen2.5-72B-Instruct: left
Qwen3-30B-A3B-Instruct: right
Llava-OneVision-72B: Left

Multiple Choice Question

Question: Which side of the person is facing the chair? (A) back (B) front (C) left (D) right

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: Where is the man positioned in relation to the woman in the image?

Answer: In front of the woman

GPT-4o: The man is positioned to the left of the woman.
GPT-4o-mini: To the left of the woman.
Claude-haiku-4.5: The man is positioned to the left of the woman.
Claude-sonnet-4.5: The man is positioned to the left of the woman in the image.
360VL-70B: Left
InternVL2_5-78B: To her left
InternVL3-78B: The man is standing to the left of the woman.
Kimi-A3B-Instruct: Standing next to her.
Qwen2.5-72B-Instruct: to the left of the woman
Qwen3-30B-A3B-Instruct: The man is positioned to the left of the woman.
Llava-OneVision-72B: Left

Multiple Choice Question

Question: Where is the man positioned in relation to the woman in the image? (A) In front of the woman (B) Behind the woman (C) To the right of the woman (D) Sitting across from the woman

Answer: A

Intern3.5-8B: A
Kimi-A3B: D
Qwen3-8B: A
Llava-OneVision-7B: D
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: Where is the woman located in relation to the man?

Answer: To his right

GPT-4o: The woman is located to the left of the man.
GPT-4o-mini: To his left
Claude-haiku-4.5: The woman is located to the left of the man.
Claude-sonnet-4.5: The woman is located to the left of the man.
360VL-70B: Left
InternVL2_5-78B: To his left
InternVL3-78B: To the left of the man.
Kimi-A3B-Instruct: Left.
Qwen2.5-72B-Instruct: to the left of the man
Qwen3-30B-A3B-Instruct: The woman is located to the left of the man.
Llava-OneVision-72B: Left

Image Input

Multiple Choice Question

Question: Where is the woman located in relation to the man? (A) To his right (B) To his left (C) Behind him (D) In front of him

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: From the perspective of the window, which side of the lady is visible?

Answer: back

GPT-4o: Left side.
GPT-4o-mini: left
Claude-haiku-4.5: Right side
Claude-sonnet-4.5: Right side
360VL-70B: Right
InternVL2_5-78B: Right
InternVL3-78B: Left side
Kimi-A3B-Instruct: Right.
Qwen2.5-72B-Instruct: left
Qwen3-30B-A3B-Instruct: left
Llava-OneVision-72B: Right

Image Input

Multiple Choice Question

Question: From the perspective of the window, which side of the lady is visible? (A) back (B) front (C) left (D) right

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: Where is the teapot located?

Answer: No, the teapot is located at the corner of the table away from the window and near the bench.

GPT-4o: On the white table near the window.
GPT-4o-mini: On the table.
Claude-haiku-4.5: On the table, on top of the wooden log/piece of wood.
Claude-sonnet-4.5: The teapot is located on the wooden table.
360VL-70B: Middle
InternVL2_5-78B: Middle
InternVL3-78B: Centered on a wooden tray.
Kimi-A3B-Instruct: The teapot is on a wooden stand in the center of the table.
Qwen2.5-72B-Instruct: The teapot is on a wooden stand in the center of the table.
Qwen3-30B-A3B-Instruct: The teapot is on a wooden stand in the center of the table.
Llava-OneVision-72B: Center

Multiple Choice Question

Question: Where is the teapot located? (A) At the corner of the table, away from the window and near the bench. (B) On the windowsill under the hanging lamps. (C) On the driftwood pieces by the window. (D) On the small round stool to the left of the table.

Image Input

Answer: A

Intern3.5-8B: B
Kimi-A3B: B
Qwen3-8B: A
Llava-OneVision-7B: B
Internvl3-78B: A

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: Describe the boy's position relative to the sheep in the photograph.

Answer: The boy is standing to the right of the sheep.

GPT-4o: Standing behind and slightly to the right of the sheep, with his hand resting on
GPT-4o-mini: He is standing behind the sheep with his hand on its back.
Claude-haiku-4.5: He is standing behind the sheep with his hand on its back.
Claude-sonnet-4.5: The boy is standing behind the sheep, holding it by its hind legs.
360VL-70B: Behind
InternVL2_5-78B: standing next to it
InternVL3-78B: Standing behind
Kimi-A3B-Instruct: The boy is standing in front of the sheep.
Qwen2.5-72B-Instruct: The boy is standing behind the sheep.
Qwen3-30B-A3B-Instruct: The boy is standing behind the sheep.
Llava-OneVision-72B: Behind

Multiple Choice Question

Question: Describe the boy's position relative to the sheep in the photograph. (A) The boy is standing to the right of the sheep. (B) The boy is standing to the left of the sheep. (C) The boy is standing behind the sheep. (D) The boy is kneeling next to the sheep.

Image Input

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: A
Internvl3-78B: A

↑ Back to List of Failure Cases

Capabilities: Location · Relation

Open-ended Question

Question: Where is the oven located in relation to the sink?

Answer: The oven is located in front of the sink.

GPT-4o: To the right of the sink.
GPT-4o-mini: To the right of the sink.
Claude-haiku-4.5: The oven is located to the right of the sink.
Claude-sonnet-4.5: To the right of the sink.
360VL-70B: Right
InternVL2_5-78B: right
InternVL3-78B: Right
Kimi-A3B-Instruct: To the right of the sink.
Qwen2.5-72B-Instruct: right
Qwen3-30B-A3B-Instruct: to the right
Llava-OneVision-72B: Right

Multiple Choice Question

Question: Where is the oven located in relation to the sink? (A) The oven is located in front of the sink. (B) The oven is behind the sink. (C) The oven is above the sink. (D) The oven is next to the trash can.

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: D
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Recognition

Open-ended Question

Question: Considering their size, how do the pizza cutter and scissors compare?

Answer: They are about the same size.

GPT-4o: I'm unable to determine the specific size comparison based on this image.
GPT-4o-mini: The pizza cutter is smaller than the scissors.
Claude-haiku-4.5: The scissors appear to be slightly larger than the pizza cutter.
Claude-sonnet-4.5: The scissors are larger than the pizza cutter.
360VL-70B: Scissors are bigger
InternVL2_5-78B: The pizza cutter is smaller than the scissors.
InternVL3-78B: The pizza cutter is longer than the scissors.
Kimi-A3B-Instruct: The pizza cutter is longer than the scissors.
Qwen2.5-72B-Instruct: The scissors are larger than the pizza cutter.
Qwen3-30B-A3B-Instruct: The scissors are larger than the pizza cutter.
Llava-OneVision-72B: The scissors are larger than the pizza cutter.

Multiple Choice Question

Question: Considering their size, how do the pizza cutter and scissors compare? (A) They are about the same size. (B) The pizza cutter is much larger. (C) The scissors are significantly bigger. (D) The pizza cutter is much smaller.

Answer: A

Intern3.5-8B: D
Kimi-A3B: C
Qwen3-8B: A
Llava-OneVision-7B: C
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Recognition

Open-ended Question

Question: Is the net in the background blue, red, or green?

Answer: None of the above

GPT-4o: Blue
GPT-4o-mini: blue
Claude-haiku-4.5: Blue
Claude-sonnet-4.5: Blue
360VL-70B: Blue
InternVL2_5-78B: blue
InternVL3-78B: Blue
Kimi-A3B-Instruct: Blue.
Qwen2.5-72B-Instruct: Blue
Qwen3-30B-A3B-Instruct: Blue
Llava-OneVision-72B: Blue

Multiple Choice Question

Question: Is the net in the background blue, red, or green? (A) None of the above (B) Blue (C) Green (D) Red

Answer: A

Intern3.5-8B: D
Kimi-A3B: D
Qwen3-8B: D
Llava-OneVision-7B: D
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases

Capabilities: Recognition

Open-ended Question

Question: What are the two boys wearing white clothes doing instead of skating?

Answer: No, they are playing skateboarding.

GPT-4o: standing
GPT-4o-mini: standing
Claude-haiku-4.5: Watching/observing the skateboarder perform the trick.
Claude-sonnet-4.5: Watching
360VL-70B: Watching
InternVL2_5-78B: jump
InternVL3-78B: Watching.
Kimi-A3B-Instruct: watching
Qwen2.5-72B-Instruct: Watching
Qwen3-30B-A3B-Instruct: Watching
Llava-OneVision-72B: Standing

Multiple Choice Question

Question: What are the two boys wearing white clothes doing instead of skating? (A) No, they are playing skateboarding. (B) Playing basketball (C) Riding bicycles (D) Playing soccer

Answer: A

Intern3.5-8B: A
Kimi-A3B: A
Qwen3-8B: A
Llava-OneVision-7B: B
Internvl3-78B: A

Image Input

↑ Back to List of Failure Cases