Do Vision Language Models infer human intention without visual perspective-taking? Towards a scalable "One-Image-Probe-All" dataset

Anonymous Author(s)
Affiliation
Address
email

Abstract

At the core of understanding the knowledge grounding of Multimodal Large Language Models (MLLMs) are two key challenges: (1) ensuring fair comparability across concepts and (2) scaling multimodal datasets to reflect real-world complexity. This paper presents a solution through the **Omni-Perspective** benchmark, which scales the construction of a 5-level question-context-answers (QCAs) **from 1 real-world image**. This benchmark pertains to 3 concepts along the Theory-of-Mind (ToM) ability hierarchy in humans and is further divided into 10 fine-grained subdifficulties. Through inference tasks, complexity, and ablation analysis, we evaluate over 2,200 consolidated QCAs on 61 MLLMs. Our findings reveal a key observation: MLLMs mostly follow the human ToM grounding pathway with exception of level-2 perspective taking. Furthermore, this dataset enables nuanced analysis of how such observations change across varying difficulty levels, modalities, distractor logic, and prompt types.

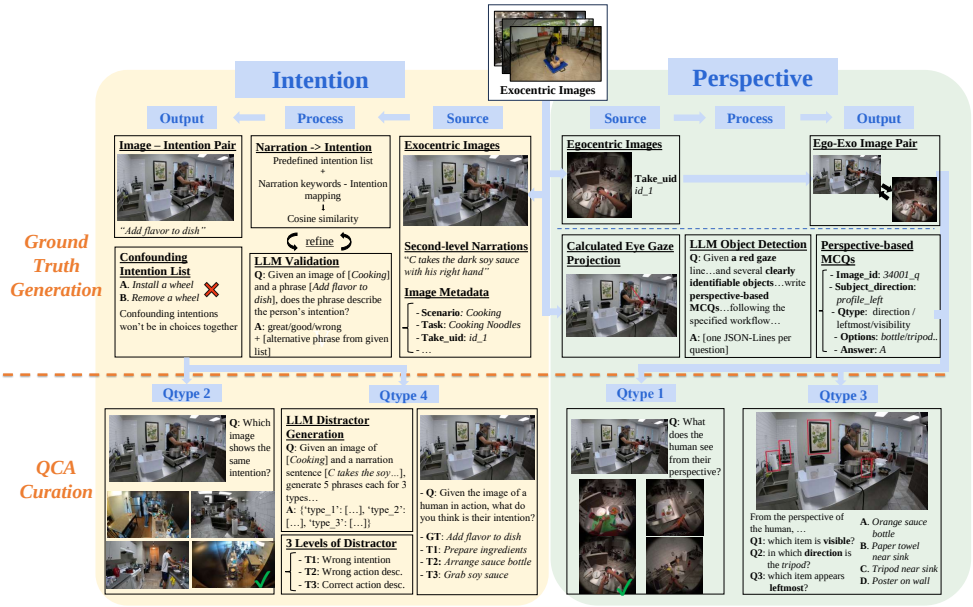

Figure 1: The scalable curation of Omni-perspective dataset

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have sparked growing interest in evaluating their capacity for complex reasoning grounded in both visual and linguistic inputs. However, rigorous assessment remains challenging due to the absence of scalable, cognitively structured benchmarks that support controlled, hierarchical, and comparative probing across diverse conceptual domains (Li et al., 2025). In this work, we address this gap by introducing a multi-image, hierarchical, and concept-controlled Question-Context-Answer (QCA) generation framework, designed to facilitate systematic evaluation of reasoning abilities across aligned tasks and cognitive levels. This framework enables the use of reusable image-intention pairs, supports fine-grained control over task difficulty, and allows for modular expansion to large-scale multimodal datasets—offering a generalizable solution for cognitively diagnostic evaluation.

A key application of this framework is the assessment of visual perspective-taking (VPT) in relation to Theory of Mind (ToM) capabilities (Premack and Woodruff, 1978; Barnes-Holmes et al., 2004; Schaafsma et al., 2015). VPT involves understanding what others see (Level 1, or VPT-1) and how they see it (Level 2, or VPT-2). Understood to be grounded in perspective-taking abilities, ToM entails modeling others' beliefs, goals, and intentions. These cognitive capacities develop in humans along a staged trajectory (Barnes-Holmes et al., 2004; Barsalou, 2008; Schurz et al., 2021), offering a natural scaffold for probing whether—and how—MLLMs internalize comparable representational structures (Sucholutsky et al., 2023).

While several benchmarks have explored vision-language reasoning, many are limited in either scope or ecological validity. For example, synthetic datasets such as CLEVR, CATER, and related benchmarks have demonstrated the utility of 3D scene modeling and controlled object manipulation for investigating compositional reasoning (Johnson et al., 2017; Girdhar and Ramanan, 2020). However, these datasets operate in highly idealized environments, characterized by clean object boundaries, minimal perceptual noise, and fully specified symbolic constraints. As a result, they tend to overestimate generalization: models trained and evaluated in these "lab-grade" settings often fail to transfer their reasoning capabilities to real-world scenes, where visual ambiguity, occlusion, temporal dynamics, and social intent are critical (Mitchell and Krakauer, 2023).

Benchmarks such as ALPRO and VQA-X expand the modality coverage and include real images or videos, but they often lack hierarchical cognitive task design or do not isolate the compositional demands of ToM-related inference. Moreover, overreliance on language priors can inflate performance in multimodal benchmarks even when visual inputs are ignored, undermining interpretability (Dongxu Li, 2022; Park et al., 2018).

To address these limitations, we propose Omni-Perspective, a cognitively motivated benchmark instantiated from our QCA generation framework. Built upon the rich, multimodal Ego-Exo4D dataset, Omni-Perspective includes over 2,200 curated QCAs structured around a six-level hierarchy that spans low-level spatial awareness to high-level belief reasoning. Each question is grounded in a shared image-intention pair and linked to a cognitive hypothesis, enabling both depth and comparability across reasoning types. Our scalable pipeline combines narration-intention mappings with GPT-4o-assisted refinement, allowing for high-quality annotation at scale without extensive manual labeling.

We evaluate 50+ MLLMs of varying modalities, sizes, and pretraining objectives, finding that while many models perform well on spatial reasoning, they falter on belief-based or intention-predictive tasks. This suggests a deviation from the developmental trajectory observed in human ToM, and motivates architectural or training-level interventions to improve grounding and inference capabilities.

In summary, this work makes three key contributions:

1. **A multi-modal probing framework** for scalable, hierarchical, and controlled Question-Context-Answer (QCA) generation, aligned with cognitive theory for systematic evaluation of multimodal reasoning.

2. **A controlled and hierarchical benchmark**, *Omni-Perspective*, designed to probe Theory of Mind (ToM) and visual perspective-taking abilities using real-world, multimodal visual data from naturalistic scenarios.

3. **An empirical analysis** revealing consistent ToM-related failure modes in state-of-the-art MLLMs, offering diagnostic insights and guiding principles for future model and training improvements.

2 Related Works

2.1 MLLM related

2.1.1 Benchark

The field of Multi-modal Large Language Models (MLLMs) requires a comprehensive evaluation of their remarkable capabilities to ensure that their development is progressing on a correct and appropriate trajectory. Early benchmarks primarily focused on single tasks, such as VQA (Antol et al., 2015), OK-VQA (Marino et al., 2019), MSCOCO (Lin et al., 2015), OCR (Liu et al., 2023), and GQA (Hudson and Manning, 2019), but have become insufficient for thoroughly assessing the broad multimodal perception and reasoning abilities of LMMs. In response, more holistic evaluations have emerged, such as LAMM (Yin et al., 2024), MM-Vet (Yu et al., 2023), SEED-Bench (Li et al., 2024), and MMBench (Liu et al., 2024c), which cover a wider range of capabilities.

2.1.2 Multi-modal Large Language Models

Recent advancements in multimodal learning have been largely driven by the unified modeling of visual and textual data using transformers (Li et al., 2019; Xu et al., 2023; Tan and Bansal, 2019; Alayrac et al., 2022; Radford et al., 2021). With the emergence of Large Language Models (LLMs), state-of-the-art (SOTA) Multi-modal Large Language Models (MLLMs) (Liu et al., 2024a; Li et al., 2023a) now integrate open-source LLMs (Touvron et al., 2023; Peng et al., 2023; Jiang et al., 2023), aligning visual features with the embedding space of LLMs (Li et al., 2023b).

To enhance open-ended conversational abilities, LLaVA (Liu et al., 2024a) introduces a method to distill the conversational capabilities of ChatGPT into MLLMs, resulting in a substantial performance boost. This approach has since become a standard procedure in the field (Wang et al., 2023; Bai et al., 2023; Gemini, 2023; Team, 2024; Sun et al., 2023; Li et al., 2022). As a result, MLLMs have demonstrated competitive performance in complex tasks requiring high-level perception and reasoning (Li et al., 2024; Liu et al., 2024a; Gemini, 2023; Fu et al., 2023; OpenAI, 2023), including spatial reasoning (Chen et al., 2024; Cai et al., 2024), character recognition (Mori et al., 1999), scene understanding (Cordts et al., 2016; Chen et al., 2017), action recognition (Jhuang et al., 2013; Herath et al., 2017), and prediction (Lan et al., 2014; Kong and Fu, 2022), often reaching near-human performance.

2.2 Visual perspective taking, Intentionality and Theory-of-Mind

The capacity to adopt another individual's visual perspective is widely recognized as a foundational component of social cognition and is considered a developmental precursor to theory of mind (ToM)—the ability to attribute mental states such as beliefs, intentions, and knowledge to oneself and others (Premack and Woodruff, 1978). While early research emphasized intention inference as central to ToM, more recent accounts have identified visual perspective taking (VPT) as a perceptual substrate supporting the emergence of mental state attribution. VPT is typically differentiated into two levels: Level-1 perspective taking (VPT-1) involves representing *what* another agent can see (i.e., which objects fall within their line of sight) whereas Level-2 perspective taking (VPT-2) entails representing *how* those objects appear from another spatial viewpoint, including their orientation and relative configuration (Kessler and Rutherford, 2010). Because VPT-2 requires mental transformations of one's egocentric reference frame—often instantiated through embodied simulation or motor imagery—it has been proposed as a particularly robust route to social understanding, even though such simulation is not strictly necessary for theory of mind reasoning in general (Hamilton et al., 2009; Gallese and Goldman, 1998; Barlassina and Gordon, 2017).

Beyond these two levels, several developmental models posit a graded trajectory in which perceptual perspective taking scaffolds increasingly abstract forms of social cognition. For example, Barnes-Holmes and colleagues propose a sequence extending from recognition of differing viewpoints to inferential use of perceptual access for epistemic judgments, prediction of actions based on true

beliefs, and ultimately the attribution of behavior based on false beliefs (Barnes-Holmes et al., 2004). Although terminological distinctions vary across frameworks, similar hierarchical structures were long proposed in traditional Piagetian theories of cognitive development (Piaget and Inhelder, 1969) and have since been elaborated in contemporary neurocognitive models that integrate perspective taking, empathy, and mental state attribution along continuous processing gradients (Schurz et al., 2021). Converging evidence from theoretical analyses suggests that tasks classified as measuring theory of mind in fact engage a distributed set of perceptual, inferential, and executive systems as opposed to being targeting a monolithic construct (Schaafsma et al., 2015; Quesque and Rossetti, 2020; Barresi and Moore, 1996). These perspectives collectively support the view that higher-order social reasoning emerges through the gradual abstraction of perceptual and embodied capacities like visual perspective taking.

This developmental progression aligns with the theoretical framework of grounded cognition, which posits that high-level cognitive functions are constitutively supported by sensorimotor systems evolved for real-world interaction (Barsalou, 2008; Gallese, 2007). Accordingly, visual perspective taking offers a principled pathway through which embodied simulation mechanisms give rise to abstract representations of others' mental states, supporting flexible and context-sensitive social inference in ecologically valid settings.

3 Omni-Perspective: A Scalable One-Image-For-All Benchmark From Visual Perspective to Intentionality Understanding

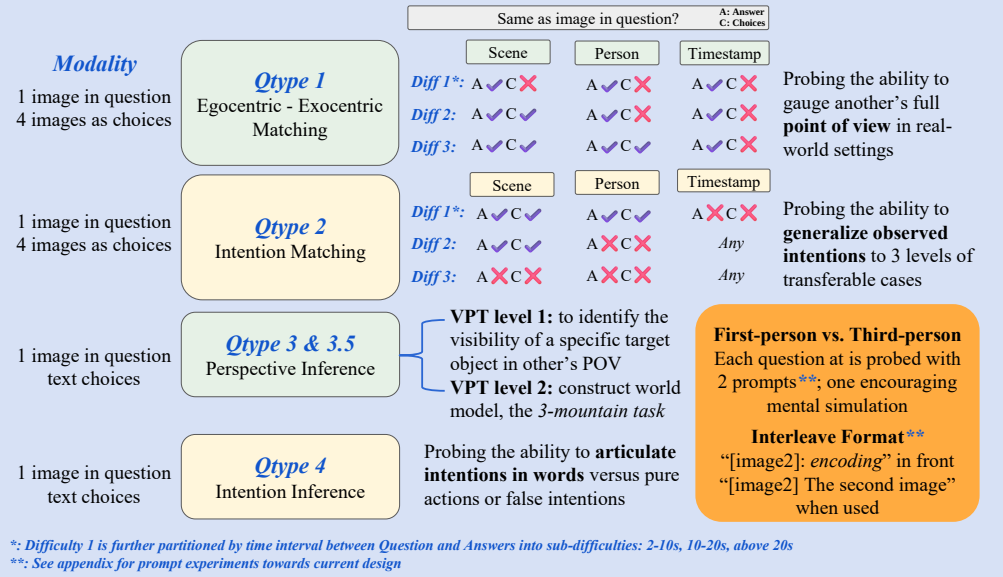

Figure 2: Overview of Omni-Perspective Bench

We define four distinct MCQ question types. Each is designed to target specific subskills aligned with the Theory-of-Mind hierarchy.

Qtype 1 (Multi-image, Egocentric - Exocentric Matching) - This question type presents the model with an exocentric image of a human in action and asks it to identify the corresponding view from four egocentric images. This task primarily probes Level-1 visual perspective-taking, requiring the model to reason about what the person sees based on spatial alignment and visual cues. Example prompt: "You are given an exocentric view of a person... Which of the following images best depicts what the person sees from their perspective?"

Qtype 2 (Multi-image, Intention Similarity) - In this task, the model is given an exocentric image of a person in action and asked to select the image depicting the most similar intention from four exocentric candidates. This question assesses the ability to generalize intention inference across individuals and scenes, contrasting with Qtype 4, which focuses on discriminating between actions

and intentions within a single context. Example prompt: "Given the image of a person performing an action... Which of the following images shows someone with a similar intention?"

Qtype 3 & 3.5 (Single-image, Spatial Perspective Inference) – The model is shown an exocentric image of a person and asked to determine the visibility or directional relation of an object from that person's perspective. All listed objects are visible in the scene, ensuring the task cannot be solved through simple object detection or visual salience heuristics. This task is inspired by the classic Piagetian "Three-Mountain Task" paradigm (Piaget and Inhelder, 1969), requiring the model to construct a Level-2 perspective-taking world model—that is, to represent not only what another agent sees, but how the scene is spatially organized from that agent's viewpoint. The model must perform an egocentric transformation of the scene, shifting reference frames to simulate another's first-person perspective. This demands an internal representation of spatial layout conditioned on agent pose and orientation. Example prompt: "From the perspective of the woman in the black shirt in the picture, which of the following items appears leftmost compared to the other choices?"

Qtype 4 (Single-image, Intention Inference) - This question presents a single exocentric image of a person in action and asks the model to choose the most likely intention from four textual options. To scale and control difficulty, distractor options are generated using a large language model (GPT-4o), conditioned on the image and atomic action annotation (See Section A.3). This format targets intention inference, requiring the model to go beyond object recognition. Example prompt: "You are given an image of a human performing an action... What do you think is their intention?"

3.1 Dataset Overview

Ego-Exo4D Dataset

We base our evaluation framework on the Ego-Exo4D dataset (Grauman et al., 2024), a large-scale, multimodal, multi-view video corpus featuring humans performing skilled activities such as cooking, bike repair, and COVID-19 self-testing. Each recording session (take) includes synchronized egocentric video from a head-mounted camera and up to four fixed exocentric views, capturing the same activity from multiple viewpoints.

The dataset is structured hierarchically across scenarios (e.g., cooking), physical settings (e.g., kitchen), takes (video sessions), cameras (time synchronized viewpoints), and annotations. Annotations include narration (atomic description of actions), procedural keysteps, and expert commentary, making it particularly suited for our use case. Our dataset includes below retrieved distribution of narrated images and goes beyond for prompt, ablation, and question evaluation analysis.

Task Type	**Total Count**
Cooking	70
Covid Test	101
Bike Repair	29
Total Tasks	$(70 + 101 + 29) \times (3 \times 2 + 2) = 200 \times 8 = 1600$

Table 1: Task counts by type

Generalization and Extensibility

Our benchmark pipeline is designed to generalize to any dataset offering (1) multi-view video and (2) action-level annotation, e.g. the LEMMA dataset (Jia et al., 2020). This modularity enables the broader application of our framework to evaluate ToM reasoning in multimodal LLMs across diverse environments and tasks.

3.2 Benchmark Overview

Scalable Ground-Truth Image-Intention Pair

We construct a scalable set of image-intention pairs that serve as the foundation for all question types in our benchmark. Four scenarios are selected based on the number of annotated takes and coverage of non-repetitive actions. For each scenario, we define a set of high-level intentions and identify representative image frames by applying a narration-keywords-to-intentions mapping. This

mapping is then empirically refined using GPT-4o, which evaluates each image-intention pair and suggests corrections when misaligned. To minimize ambiguity, intentions that are visually similar (e.g. *install a wheel* and *remove a wheel*) or sequentially entailed (e.g. *set up test* and *perform test*) — referred to as confounding distractors — are excluded from co-occurrence within the same question. This iterative process enables scalable generation of high-quality ground-truth image-intention pairs. Refer to Section A.1 for more technical details.

Comparability across Question Types

Reusing images across question types - Each image-intention pair links to both egocentric and exocentric views that are time-synchronized within the same take. This allows the same visual context to be used for both perspective and intention questions, minimizing variability arising from differences in scene content.

Consistent question phrasing - We standardize the linguistic structure of prompts across all question types, avoiding shortcut through language cues. This reduces the risk of models exploiting superficial lexical patterns and promotes a fairer assessment of reasoning capabilities.

Uniform image abstraction level - All images are sampled from real-world video footage with similar resolution, camera specification, and background complexity. This avoids confounding effects associated with abstraction level — such as those seen when mixing synthetic, staged, or cartoon images with natural scenes — and ensures that all questions have perceptually comparable visual input.

First- and Third-Person Language Query

Each question type is presented in both first-person and third-person point-of-view to distinguish between two levels of perspective-taking. First-person prompts (e.g., "If you were the person in the image, what is in your line of sight?") encourage the model to take the subject's role, reflecting a mental simulation of world model and thus Theory-of-Mind reasoning (Barresi and Moore, 1996). Third-person prompts (e.g., "Given the image with a person in action, what is their intention?") treat the model as an external observer, targeting Level-1 perspective-taking.

Distractors with Multiple Difficulty Levels or Types

Qtype 1 and 2 in our benchmark are presented at three levels of difficulty, defined by the design of distractor choices. Difficulty increases as distractors become visually similar to the correct answer (e.g. comparable objects or spatial arrangements), while easier distractors differ more clearly in object type or environment setting. Qtype 4 does not use fixed difficulty levels but instead includes three semantically distinct distractor types, ranging from low-level action descriptions to high-level intentions. This controlled variation allows us to probe the robustness and granularity of model reasoning under varying cognitive demand.

4 Experiment

4.1 Setup

Inference: With the curated QCA-prompt, we assessed an extensive collection of models spanning a wide spectrum of architectures, parameter scales, and training methodologies. Our study encompassed a total of 61 MLLMs. The selection included prominent proprietary models such as those from the ChatGPT and Claude families, chosen for their established performance and widespread use. The open-source cohort featured state-of-the-art models, including InternVL, the Qwen series, and the recently released DeepSeek models, which have received increasing attention for their strong performance in multimodal tasks. The open-source models under evaluation ranged in size from 1 billion to 110 billion parameters, enabling detailed performance analysis across scales. Proprietary models were evaluated through API calls on standard personal computers. For open-source models, we performed inference locally on a compute cluster equipped with 8×NVIDIA A100 80GB GPUs. In practice, models under 13B parameters were typically executed on a single GPU, models between 13B and 32B required two GPUs, those between 32B and 70B utilized four GPUs, and models exceeding 70B ran across all eight GPUs. We adhered closely to the official inference codebases provided by model developers to ensure reproducibility and preserve model-specific inference optimizations. To further ensure consistency and correctness in handling multimodal inputs, we developed a unified evaluation toolkit capable of parsing and validating model responses across varying input formats.

Evaluation: To determine correctness, the model's selected option is compared against the ground truth, with any instance labeled as FAIL in the matching process automatically marked incorrect. Specifically: 1) Template aatching is attempted first, using a set of pre-defined output formats to map the model's response to one of the answer choices. 2) If template matching fails, the instance is passed to LLM matching, where a large language model—Llama-3.1-70B-Instruct(Grattafiori et al., 2024)—acts as a semantic judge to infer the intended answer choice.

To reduce the influence of answer-position bias, we adopt circular evaluation (Liu et al., 2024b). In this method, the multiple-choice options for each question are rotated across all possible positions. The model must correctly answer all k permutations of a k-choice question to be considered accurate—ensuring that its success is not due to token position or randomness.

4.2 Main Results

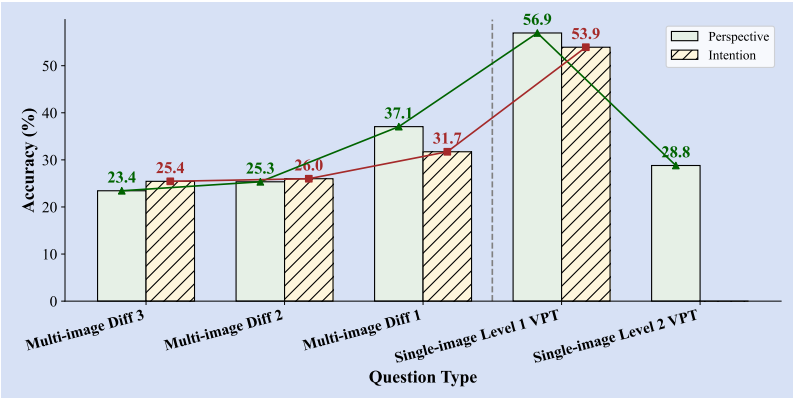

Figure 3: Comparative result between perspective taking and intention understanding across different difficulty levels and input types.

Visual Perspective Grounding in Multi-Modal Large Language Models We present comparative results (perspective vs. intention) across different difficulty levels (difficulty 1, 2 and 3) and input settings (single v.s. multi-image) in Figure 3. Several expected observations validate our benchmark design: 1. As difficulty increases from left to right (in the left section of the dashed line), both perspective and intention performance improve. 2. Performance on single-image tasks is consistently higher than on the three levels of multi-image tasks (to the right vs. left of the dashed line), largely due to the limited ability of MLLMs to process multi-image inputs.

Surprisingly, except for difficulty-3, where perspective is on par with intention, all other comparisons (difficulty-2, difficulty-1, and single-image) show better performance in perspective taking than in intention understanding. This contrasts with prior work Gao et al. (2025); Li et al. (2025). To further explore this distinction, we evaluate performance on level-2 perspective taking, specifically the three-mountain task (rightmost bar in Figure 3). In a fair comparison (both single-image), the three-mountain task performs lower than intention understanding, which aligns with previous findings Gao et al. (2025); Li et al. (2025). This suggests that the discrepancy between intention and level-2 perspective taking is not due to a lack of visual perspective-taking ability, but rather factors such as limited spatial reasoning in the current MLLMs.

Does prompting for Mental Simulation help? Encouraging mental simulation (putting oneself in another's shoes) is discussed to potentially benefit both visual perspective taking and intention understanding ability, raising an intriguing question: Does explicitly prompting MLLMs to perform mental simulation improve performance on these tasks (Barlassina and Gordon, 2017)? A drill down into single image-prompt pairs (less confounded by distractor selection methods) shows that prompting MLLMs with first-person phrasing significantly improves performance on perspective-taking tasks ($p = 0.0321$) on spatial reasoning, while remaining inconclusive for intention understanding.

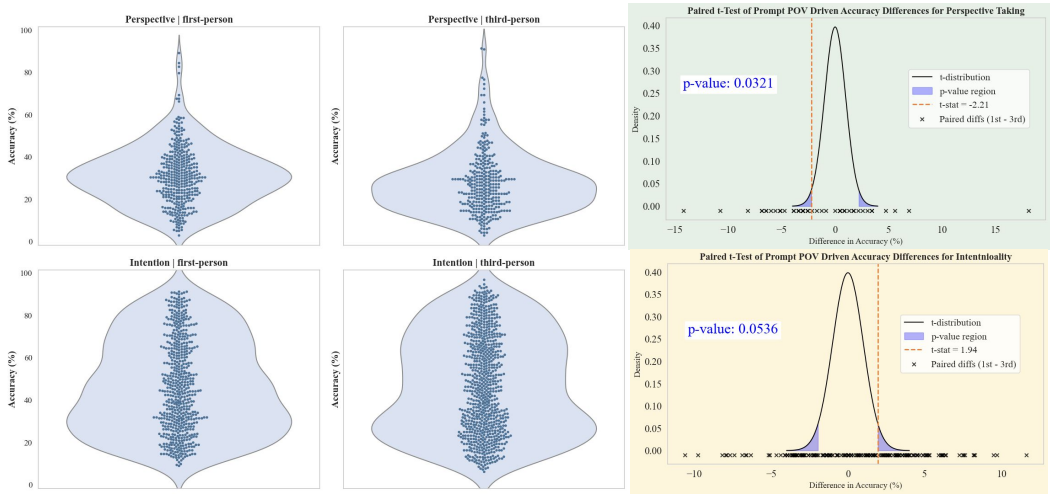

Figure 4: **Left:** Distribution of accuracy partitioned by probing concept and point-of-view of prompt; **Right:** Paired-T test results of single-image question for 2 types of prompts

4.3 Distractor Ablation Tests

For Qtype 4 - where distractors differ semantically (e.g. action descriptions versus high-level intentions) - we randomly select and mix choices from all three types for 200 questions. We then construct an additional ablation set of 95 randomly selected questions, each replicated into three versions containing distractors exclusively from one type. All other variables, including the image, prompt wording, and correct answer, remain constant for controlled comparison.

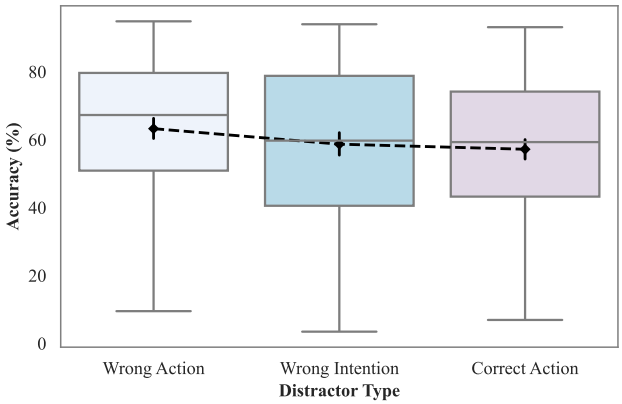

Figure 5: Accuracy by distractor type in Qtype 4 Ablation Test where the distractor type is controlled

Figure 5 reveals that average model accuracy varies across distractor types. Compared to the original Qtype 4 setup with an average accuracy of 53.9% (Figure 3), the ablation set yields consistently higher performance. This improvement likely stems from the reduced semantic variability, allowing models to exploit language-based shortcuts. Among the distractor types, wrong action results in the highest accuracy, which may be attributed to its double-layered deviation from the correct answer: it involves low-level action or object recognition rather than high-level intention inference, and the action described is itself incorrect, limiting the model's ability to rely on object-centric heuristics.

4.4 Benchmark Results

Stronger Models Exhibit Greater Differentiability on Easier Tasks Accuracy varies widely across models at lower difficulty levels, with top-performing models such as *llava-video-72b-*

Model	Qtype 1 *Ego-Exo Match*			Qtype 2 *Intention Match*			Qtype 3 *Perspective Inference*	Qtype 4 *Intention Inference*
	Diff1	Diff2	Diff3	Diff1	Diff2	Diff3		
GPT-4o	97.24%	**46.09%**	28.09%	75.87%	36.28%	30.60%	31.37%	59.35%
deepseek-vl2-small	40.57%	41.98%	**41.36%**	71.81%	**73.47%**	**75.93%**	**57.08%**	43.45%
Qwen2.5-VL-72B-Instruct	95.99%	45.05%	35.75%	**79.26%**	34.95%	32.41%	41.27%	61.38%
LLaVA-Video-72B-Qwen2_multi_frame	**98.35%**	42.69%	29.91%	68.62%	35.20%	37.96%	46.93%	59.23%
LLaVA-Video-7B-Qwen2_multi_frame	95.28%	38.44%	17.99%	67.55%	35.46%	41.67%	48.11%	51.73%
VILA1.5-40b	96.46%	32.78%	31.78%	56.91%	29.34%	23.15%	35.38%	**75.68%**
Mantis-8B-Idefics2	75.88%	39.25%	28.39%	66.57%	37.18%	32.33%	32.08%	59.85%
Llama-3-LongVILA-8B-256Frames	26.18%	29.72%	26.87%	59.04%	58.67%	58.33%	35.14%	73.88%
llava_next_interleave_7b	67.25%	26.55%	21.73%	49.71%	27.56%	26.72%	34.20%	64.38%
Llama-3-VILA1.5-8B	72.17%	28.30%	21.96%	40.43%	23.72%	23.15%	35.38%	60.93%
Ovis1.6-Gemma2-9B	69.50%	30.44%	25.88%	31.10%	25.64%	28.45%	44.34%	46.15%
Janus-Pro-1B	24.76%	26.18%	25.23%	43.09%	52.55%	56.48%	23.82%	32.50%
Vintern-3B-beta	44.88%	24.48%	25.88%	30.23%	25.51%	26.29%	35.38%	57.45%
InternVL2-4B	28.38%	24.09%	26.63%	37.79%	24.36%	23.71%	41.75%	51.00%

Table 2: Accuracy by model on each Qtype subtask. Best cells are bold and both best and second-best are shaded.

qwen2_multi_frame achieving near-perfect scores (98% on Qtype 1 Difficulty 1), while many others remain below 30%. This variance diminishes as task difficulty increases: the standard deviation in accuracy drops from nearly 20% at Difficulty 1 to under 5% at Difficulty 3. This pattern is most evident among stronger, higher-capacity models, which show clear separation on simpler tasks but converge to similarly low accuracy as complexity rises. Weaker models, by contrast, perform consistently poorly across all levels with limited differentiation.

Model Series Show Consistent Performance Trends Certain model series consistently outperform others. The *qwen2_5_vl_series* and *llava_video_multiframe_series* perform especially well at larger scales, often scoring above 50% across tasks. Conversely, the *eagle_series_x4* and *x5* models underperform broadly; even the 13B variant *eagle-x4-13b-plus* averages below 20%, suggesting potential limitations in architecture, pretraining, or fine-tuning strategies.

Scaling Model Size Yields Diminishing Returns Beyond a Point Larger models generally outperform their smaller counterparts. For instance, in the *vila_series*, *vila1.5-40b* achieves a mean accuracy of 48%, outperforming *vila1.5-13b* (39%) and *vila1.5-3b* (33%). However, some series show marginal benefits from scaling: *llava-video-72b-qwen2_multi_frame* only slightly outperforms its 7B counterpart (52% vs. 50%), and within *internvl2_series*, the jump from 2B to 40B offers limited accuracy improvement. This suggests that beyond a certain threshold, increases in model size alone may not yield proportionate gains.

5 Discussion

This study introduces the Omni-Perspective benchmark, a cognitively grounded and scalable framework for probing MLLMs along the developmental hierarchy of ToM reasoning. We find that while models perform reliably on Level-1 perspective-taking tasks, they consistently struggle with Level-2 visual perspective-taking and intention inference. This pattern generally aligns with developmental theories suggesting that higher-order social reasoning builds upon more basic perceptual capacities, and is thus inherently more demanding. This suggests that MLLMs may be situated within a human-like developmental trajectory for social cognition, albeit currently limited to lower levels of the hierarchy. The observed performance gap reveals a key limitation in current MLLMs: their limited capacity for mental simulation—a mechanism believed to support flexible, context-sensitive social inference. Furthermore, our ablation studies show that model behavior is highly sensitive to distractor configurations and prompt phrasing, indicating a reliance on superficial cues rather than robust mental state representations. Taken together, the Omni-Perspective benchmark offers a controlled and interpretable framework for evaluating social reasoning in MLLMs, while also providing diagnostic insights into their architectural and training limitations.

In the meantime, we acknowledge that our benchmark relies on videos with sustained, non-transient task focus as a proxy for intentionality, which may not generalize to brief or socially nuanced intentions. It also assumes access to multiple viewpoints, limiting applicability to monocular settings.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Barlassina, L. and Gordon, R. M. (2017). Folk psychology as mental simulation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.

Barnes-Holmes, Y., McHugh, L., and Barnes-Holmes, D. (2004). Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today*, 5(1):15–25.

Barresi, J. and Moore, C. (1996). Intentional relations and social understanding. *Behavioral and brain sciences*, 19(1):107–122.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

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

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., and Xia, F. (2024). Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

Dongxu Li, Junnan Li, H. L. J. C. N. S. C. H. (2022). Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., and Ji, R. (2023). Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*.

Gallese, V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):659–669.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501.

Gao, Q., Li, Y., Lyu, H., Sun, H., Luo, D., and Deng, H. (2025). Vision language models see what you want but not what you see.

Gemini (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*.

Girdhar, R. and Ramanan, D. (2020). CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy,

S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The llama 3 herd of models.

Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.-J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M. M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K. J., Liu, J.-W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S. K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S. C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G. M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C. V., Newcombe, R., Park, H. S., Rehg, J. M., Sato, Y., Savva, M., Shi, J., Shou, M. Z., and Wray, M. (2024). Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives.

Hamilton, A. F. d. C., Brindley, R., and Frith, U. (2009). Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition*, 113(1):37–44.

Herath, S., Harandi, M., and Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21.

Hudson, D. A. and Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M. J. (2013). Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199.

Jia, B., Chen, Y., Huang, S., Zhu, Y., and chun Zhu, S. (2020). Lemma: A multi-view dataset for learning multi-agent multi-task activities.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. *arXiv preprint arXiv: 2310.06825*.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Kessler, K. and Rutherford, H. (2010). The two forms of visuo-spatial perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in Psychology*, Volume 1 - 2010.

Kong, Y. and Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401.

Lan, T., Chen, T.-C., and Savarese, S. (2014). A hierarchical representation for future action prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 689–704. Springer.

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. (2024). Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., and Si, L. (2022). mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv: 2205.12005*.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023a). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CONFERENCE*.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023b). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, Y., Gao, Q., Zhao, T., Wang, B., Sun, H., Lyu, H., Luo, D., and Deng, H. (2025). Core knowledge deficits in multi-modal language models.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2024a). Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., and Lin, D. (2024b). Mmbench: Is your multi-modal model an all-around player?

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. (2024c). Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Liu, Y., Li, Z., Yang, B., Li, C., Yin, X., Liu, C.-l., Jin, L., and Bai, X. (2023). On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Mori, S., Nishida, H., and Yamada, H. (1999). *Optical character recognition*. John Wiley & Sons, Inc.

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*.

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. (2023). Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Piaget, J. and Inhelder, B. (1969). *The Psychology of the Child*. Basic Books, New York.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Quesque, F. and Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv: 2103.00020*.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., and Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in cognitive sciences*, 19(2):65–72.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., and Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological bulletin*, 147(3):293.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., et al. (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. (2023). Generative multimodal models are in-context learners. *Computer Vision and Pattern Recognition*.

Tan, H. and Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Team, C. (2024). Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. (2023). Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., and Duan, N. (2023). Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647.

Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al. (2024). Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. (2023). Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Appendices

A Dataset Details

A.1 Ground-Truth Image-Intention Pair Generation

The section contains the essential information used to scale the ground-truth image-intention pair generation process. Below, we detail key design choices and procedures.

Scenario and Task Selection - Scenarios and tasks with repetitive behaviors (e.g., dancing, instruments playing) are excluded. The Table 3 lists all scenarios and tasks considered.

Table 3: Scenario and Applicable Tasks

Scenario	Applicable_task_name
Bike Repair	Install a Wheel, Remove a Wheel, Fix a Flat Tire - Replace a Bike Tube, Clean and Lubricate the Chain
CPR	First Aid - CPR
Covid Test	Covid-19 Rapid Antigen Test
Cooking	Making Cucumber & Tomato Salad, Making Greek Salad, Making Sesame-Ginger Asian Salad, Making Chai Tea, Making a Milk Tea, Cooking Noodles, Cooking an Omelet, Cooking Scrambled Eggs, Cooking Tomato & Eggs, Cooking Dumplings, Cooking Pasta, Cooking Sushi Rolls, Cooking Samosas, Making Greek Salad, Making White Radish & Lettuce & Tomato & Cucumber Salad

Intention Definition and Keywords Mapping - For each selected scenario, we define a set of high-level intentions (Table 4). We apply a two-stage matching process:

1. For each take, we extract all action-level narrations and compute cosine similarity between narration sentences and the keyword list associated with each intention (Table 5).
2. From each take, we select up to three frames (from the annotated *best_exo* camera) with the highest similarity scores for each intention, ensuring a minimum 10-second separation to avoid look-alike images. These are used as first-pass image-intention candidates.

Table 4: Scenarios and Associated Intentions

Scenario	Intention
Bike Repair	Install a wheel
	Replace the tire tube on the wheel
	Clean and lubricate the chain
	Remove a wheel
CPR	Confirm patient consciousness
	Call for help
	Press for heart rate
Covid Test	Set up for test
	Understand instruction
	Perform test
Cooking	Prepare ingredient
	Preheat pan for cooking
	Add flavor to dish
	Clean up work station

Table 5: Intention to Keywords Mapping

Intention	Keywords
Install a wheel	install, attach, bike fork
Replace the tire tube on the wheel	tire level, tire valve, inflate/deflate, tire tube, bike inner tube, fit the bike tire
Clean and lubricate the chain	chain lube, degreaser spray, lubricant bottle, hold the towel, clean the chain, pick up a brush, spray water
Remove a wheel	removes the bicycle wheel, removes the wheel, take off wheel
Confirm patient consciousness	pat, check for breathing, observe, tap
Call for help	wave her hands, extend right hand, extend left hand, call for help
Press for heart rate	interlace the fingers of this hands, compress, interlock, press
Set up for test	put on desk, place on desk, pick out from box, set up, open the box
Understand instruction	test manual, test instruction, read, understand, flip
Perform test	insert test swab, pick up the collection swab, dip the swab, nostril, nose
Prepare ingredient	chopping board, tomato, onion, scallion, knife, cut, carrot, potato, banana
Heat pan for cooking	press a switch, take the skillet, turn on heat, adjust the heat, turn on gas stove, picks the frying pan
Add flavor to dish	pick up black pepper, pick up the salt, soy sauce, sauce, sugar
Clean up work station	wash, turns on the tap, opens the tap, waste bins, push dirt into sink hole, picks the dirt, trash can

Confounding Distractors - As shown in Table 6, for some intentions, we define the confounding distractors that are either visually similar with or sequentially entailed to each other, and avoid presenting them within the same question.

Table 6: Intention and Confounding Distractor Pairs

Intention	Confounding Distractor
Install a wheel	Remove a wheel
Remove a wheel	Install a wheel
Confirm patient consciousness	Press for heart rate
Press for heart rate	Confirm patient consciousness
Set up for test	Perform test
Understand instruction	Set up for test
Perform test	Set up for test
Prepare ingredient	Clean up work station
Clean up work station	Prepare ingredient

LLM Validation - We then use GPT-4o to validate each image-intention pair.

Sample Prompt:

Figure 6: Sample Image Input for LLM Qtype4 Distractor Generation - Cooking

- I will provide an image of a person performing an action related to *Cooking* (*note: Scenario*), and a phrase that tries to describe the intention of the person: *"Add flavor to dish"* (*note: Intention*). Return only the required strings in a list format based on the following instructions, without additional explanations.
- Return 'great' if you are confident that the phrase accurately describes the intention of the person in the image.
- Return 'good' if you think the phrase describes the intention, but not as confidently.
- Return 'wrong' if the phrase is unrelated to the image, is not the intention that a normal non-technical human viewer could infer from the image, or has a better alternative from the following list: *[Prepare ingredients, Clean up work station, Add flavor to dish, Preheat pan for cooking]* (*note: All intentions in the scenario*).
- If you choose 'wrong', also return the best alternative option from the list. If none of the alternatives work, return 'None'.

A.2 Qtype 3 Question Generation

We utilize GPT-3o to scale the question generation process for Qtype3. Below documents the detailed prompt we provide to the LLM.

Context

You will receive one or more third-person photos of everyday scenes. Each image contains:

1. a **red gaze line** that starts at the eyes of the **primary person** (the "subject"), and
2. several clearly identifiable objects.

Your task is to write **perspective-based multiple-choice questions (MCQs)** that test spatial reasoning **from the subject's viewpoint** (not the camera's).

MCQ Templates

- Type: Visibility - From the perspective of SUBJECT, which of the following items in the image are visible?
- Type: Direction - From the perspective of SUBJECT, in which direction is TARGET-OBJECT?
- Type: Leftmost/Rightmost - From the perspective of SUBJECT, which of the following items appears leftmost / rightmost?

Note on choices: All options must be generic and unambiguous (e.g., "a red box on the counter" rather than "a toolbox"). Label the correct answer A–D.

Workflow

1. **Load the image**
   (a) Note the general setting (kitchen, bike workshop, etc.).
   (b) Locate the subject (person with the red line).
   (c) **Determine subject orientation** — choose exactly one:
      - facing-camera
      - back-to-camera
      - profile-left (subject looking toward **camera-left**)
      - profile-right (subject looking toward **camera-right**)

   If the body is roughly 45°, combine them, such as facing-camera & profile-right
   (d) **Build a subject-centric frame**
      - **Forward** = the red gaze line.
      - **Left / Right** = rotate the frame ± 90° around the subject.

Subject Orientation	Subject-Left	Subject-Right	Quick Visual Cue
facing-camera	camera-**right**	camera-**left**	(mirror rule)
back-to-camera	camera-**left**	camera-**right**	(mirror rule)
profile-left	**down** in photo	**up** in photo	
profile-right	**up** in photo	**down** in photo	

      - **Behind** = opposite of forward.
      - If subject orientation is combined (e.g., facing-camera & profile-right), the projection should also be combined.
2. **Parse objects**
   List every salient object as *minimal-adjective + generic noun* (e.g., "blue mug," "metal faucet"). Re-use these exact names in the MCQs.
3. **Generate three MCQs (one of each type) per image**
   - Describe the subject succinctly (e.g., "the woman in a blue apron").
   - **Direction**: pick a clear {TARGET-OBJECT}; options = front / behind / left / right.
   - **Visibility & Leftmost/Rightmost**: provide four distinct objects.
   - Mark the correct answer.
4. **Quality check (mandatory)**
   - Verify every spatial relation in the subject-centric frame.
   - Ensure wording is concise, bias-free, and each referenced object is clearly visible.
5. **Output** — one `JSON` record per question. {
   "image_id": "<image filename or UID>",
   "subject_direction": "facing-camera | back-to-camera | profile-left | profile-right | <combined>",
   "question_type": "visibility | direction | leftmost | rightmost",
   "question": "<full question text>",
   "options": "A": "...", "B": "...", "C": "...", "D": "..." ,
   "answer_key": "A/B/C/D"
   }

A.3 Qtype 4 Distractor Generation

The distractor generation process for Qtype 4 requires special attention due to its textual nature.

For **Wrong Intention** distractor type, we randomly sample other intentions from the same scenario, while explicitly avoiding confounding distractors (Table 6). When the number of suitable alternatives is insufficient, we supplement the set with manually created pseudo-intentions that are plausible yet not part of our dataset (e.g. *Taste the food*, *Throw away food waste*).

For **Wrong Action** and **Correct Action** distractor types, we leverage a LLM (GPT-4o) to scale generation and validation.

Sample Prompt:

Figure 7: Sample Image Input for LLM Ground-Truth Validation - Cooking

You are an expert in linguistics and are good at coming up natural alternative expression if given a sentence in English.

Give the sentence '*C takes the dark soy sauce with his right hand.*', please come up with the following, without including any explanations.

1. Type 3: 5 concise phrases that describe the action (atomic description) in the sentence. If the sentence doesn't have 'C' (a human) as the subject, make sure to phrase the action such that it sounds reasonable if the subject is a human.

2. Type 2: 5 concise phrases that describe different but similar actions. For example, these alternate phrases can EITHER a) describe the same action on a different object, OR b) describe different action on the same object. Do not replace both action and object at the same time. It is preferred that if a human is to perform these phrases, their body gestures and/or scenario will look like the original sentence.

General requests:

1. return phrases without explicit subject. For example, 'C does something' should be shortened to 'do something'.
2. the phrases should use verbs and nouns that are natural and colloquial.
3. the phrases should make sense with human as the subject, even if the subject in original sentence may not be a human. Rephrase the original sentence to human-subject first, then generate alternatives.

The output format should follow: {'type_3': [phrases1, phrases2, ...], 'type_2': [phrases1, phrases2, ...]}

Sample Output:

{'type_3': ['grab soy sauce', 'hold dark soy', 'pick up sauce', 'lift dark soy', 'take soy bottle'], 'type_2': ['grab light soy sauce', 'hold ketchup bottle', 'pick up olive oil', 'lift sesame oil', 'take vinegar bottle']}

B NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both the abstract and introduction crystallize the curation structure, use case, and scaling goal of our benchmark with literature, statistics, and procedural visualizations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

Justification: The paper discusses several limitations, including reliance on synthetic prompts that may not generalize across all domains, limited access to closed-source models for full comparison, and the challenge of verifying whether LLMs perform true mental simulation or merely pattern match.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: The paper does not present theoretical results or formal proofs. It is primarily an empirical benchmark contribution.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: The paper includes detailed descriptions of the dataset, the benchmark structure, prompt generation pipelines, and evaluation procedures. All steps required to reproduce the experimental setup are clearly documented in both the main paper and appendix

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset and code will be made available upon the sharing of private link. Instructions for dataset usage and evaluation are provided in paper body and in the supplemental material. The paper also includes a structured overview of asset preparation. Further annotation and usage instructions will be added upon official release to the public.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all relevant experimental configurations, including prompt types, number of frames per input, prompting strategies, and accuracy calculation. Full prompting templates are included in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We performed statistical significance testing and showed the results in 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We explained our experiment setup and inference details in Section 4.1

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We made sure to conform with the Code of Ethics. Please see the below bullet points for more explanation.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper presents a foundational benchmark for evaluating multimodal reasoning in machine learning models. It does not propose a deployed system or application, nor does it yet involve real-world users or decision-making contexts. As such, the work does not present direct societal impacts—positive or negative—in its current form.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The constructed dataset contains only predefined and reviewed output, and thus poses no such risks,

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credited and introduced the dataset we use in the paper in the Section 3.1 and in References. All authors have been granted licenses to download and use the dataset from the official website.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The benchmark dataset and code will be made available upon the sharing of private link. Instructions for dataset usage and evaluation are provided in paper body and in the supplemental material. The paper also includes a structured overview of asset preparation. Further annotation and usage instructions will be added upon official release to the public.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use LLMs for generating our VQA questions. We documented the detailed procedures and prompts in Sections A.1, A.2, and A.3

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.