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

Bingyang Wang [1] **Yijiang Li** [2] **Qingyang Zhou** [*3] **Hui Yi Leong** [*4] **Tianwei Zhao** [*5] **Letian Ye** [6] **Hokin Deng** [6] **Dezhi Luo** [7] **Nuno Vasconcelos** [2]

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.

1. Introduction and Related Works

The rapid development of Multi-modal Large Language Models (MLLMs) necessitates robust benchmarks to evaluate their reasoning capabilities. Early evaluations targeted specific tasks such as VQA (Antol et al., 2015), OK-VQA (Marino et al., 2019), MSCOCO (Lin et al., 2015), and GQA (Hudson & Manning, 2019), but these are insufficient for assessing the broader cognitive and perceptual abilities of modern MLLMs. Recent benchmarks like LAMM (Yin et al., 2024), MM-Vet (Yu et al., 2023), SEED-Bench (Li et al., 2024), and MMBench (Liu et al., 2024) offer wider task coverage, yet they often lack hierarchical design or cognitively grounded task structures. Synthetic datasets such as CLEVR and CATER enable controlled investigations of compositional reasoning (Johnson et al., 2017; Girdhar & Ramanan, 2020), but their idealized environments limit generalizability to real-world scenes, where ambiguity, occlusion, and social cues are critical (Mitchell & Krakauer, 2023). Datasets like ALPRO and VQA-X expand to real images or videos, but typically do not isolate Theory-of-Mind (ToM)-related reasoning or scaffold tasks across cognitive levels (Dongxu Li, 2022; Park et al., 2018).

A central application of cognitively structured benchmarks is evaluating visual perspective-taking (VPT) and its connection to ToM—the capacity to attribute beliefs, intentions, and knowledge to others (Premack & Woodruff, 1978; Barnes-Holmes et al., 2004). VPT-1 refers to knowing what another agent can see; VPT-2 requires inferring how things appear from that agent's viewpoint, often involving mental spatial transformations (Kessler & Rutherford, 2010; Hamilton et al., 2009). These capacities support social cognition and form the foundation for more abstract ToM reasoning (Gallese & Goldman, 1998; Barlassina & Gordon, 2017).

Developmental models describe a gradual trajectory from simple visual access recognition to belief attribution, including true and false beliefs (Barnes-Holmes et al., 2004; Schurz et al., 2021; Piaget & Inhelder, 1969). Grounded cognition theories argue that high-level social reasoning emerges from perceptual and sensorimotor processes evolved for real-world interaction (Barsalou, 2008; Gallese, 2007). This supports the idea that perspective taking serves as a scaffold for inferring mental states in ecologically complex situations.

We introduce Omni-Perspective 1, a cognitively grounded benchmark instantiated via a scalable, hierarchical QCA (Question-Context-Answer) generation framework. Built from the multimodal Ego-Exo4D dataset, it includes 2,200+ curated QCAs across six reasoning levels—from spatial visibility to belief-based inference. Each question is tied to a shared image-intention pair and a specific cognitive construct, enabling both within- and across-task comparability.

[*]Equal contribution [1]Emory University [2]University of California, San Diego [3]Northwestern University [4]University of Chicago [5]Johns Hopkins University [6]Carnegie Mellon University [7]University of Michigan, Ann Arbor. Correspondence to: Bingyang Wang <wangby.icy@gmail.com>, Yijiang Li <yijiangli@ucsd.edu>.

ICML 2025 Workshop on Assessing World Models.

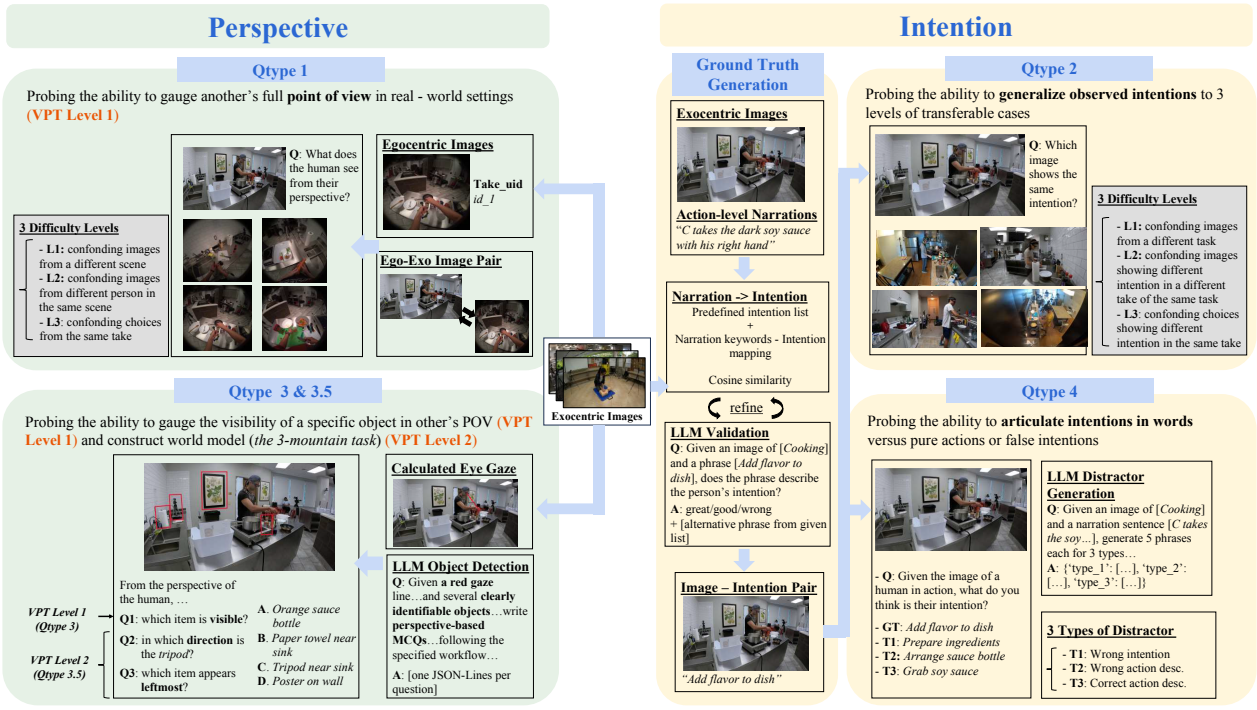

Figure 1: The scalable curation of Omni-perspective dataset

Using a pipeline that combines metadata with LLM-assisted refinement into QCAs, we generate high-quality annotations with minimal manual overhead. Evaluation across 61 MLLMs shows strong spatial reasoning but limited ToM-related inference, revealing a divergence from human developmental patterns and underscoring the need for more grounded, cognitively structured training approaches.

2. Omni-Perspective: A Scalable Benchmark From Perspective-taking to Intentionality

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) - The model is given an exocentric image of a person in action and must identify the corresponding egocentric view from four options. This probes Level-1 visual perspective-taking, requiring reasoning about spatial alignment and visual cues. Example: "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) - The model is shown an exocentric image of a person and must select the image depicting the most similar intention from four exocentric options. This tests the ability to generalize intention inference across scene. Example: "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 must determine the spatial relationship or visibility of objects from that person's perspective. While visibility question (Qtype 3) test Level-1 perspective-taking, Qtype 3.5 requires 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. Example: "From the perspective of the woman in the black shirt in the picture, which of the following items appears leftmost compared to the others?"

Qtype 4 (Single-image, Intention Inference) - The model is given an exocentric image of a person and must choose the most likely intention from four textual options. 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 and to discriminate between actions and intentions within one context. Example: "Given an image of a human performing an action... What do you think is their intention?"

2.1. Dataset Overview

Our base dataset—Ego-Exo4D—is a large-scale, multimodal, multi-view video corpus capturing skilled human activities like cooking, bike repair, and COVID-19 self-testing (Grauman et al., 2024). Each session includes synchronized egocentric video from a head-mounted camera and up to four fixed exocentric views, providing comprehensive multi-view coverage of the same activity. The dataset is hierarchically organized by scenarios (e.g., cooking), physical settings (e.g., kitchen), sessions, cameras, and annotations. Rich annotations, including narrated action descriptions, procedural keysteps, and expert commentary, make it ideal for evaluating Theory-of-Mind (ToM) reasoning in multimodal language models. Our benchmark pipeline, leveraging this dataset, is designed to generalize to any multi-view, action-annotated dataset such as LEMMA (Jia et al., 2020), enabling extensible evaluation of ToM reasoning across diverse environments and tasks.

2.2. Benchmark Overview

We construct a scalable set of image-intention pairs 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. High-level intentions are defined, and representative image frames are identified using a narration-keywords-to-intentions mapping, refined with GPT-4o to correct misalignments. Confounding distractors, such as visually similar intentions (e.g., *install a wheel* vs. *remove a wheel*) or sequentially entailed actions (e.g., *set up test* vs. *perform test*), are excluded to minimize ambiguity. This ensures scalable, high-quality ground-truth image-intention pairs (see Section A.1 for details).

- **Reusing images across question types**: Each image-intention pair links to both time-synchronized egocentric and exocentric views, allowing consistent visual context across perspective and intention questions.
- **Consistent question phrasing**: Prompts are standardized to avoid linguistic shortcuts and ensure fair assessment of reasoning capabilities.
- **Uniform image abstraction level**: Images are sampled from real-world video footage with consistent resolution, camera specifications, and background complexity, avoiding confounding effects from mixing synthetic or staged images with natural scenes.
- **First- and third-person language queries**: Questions are presented in both first-person (e.g., "If you were the person in the image, what is in your line of sight?") and third-person (e.g., "Given the image, what is the person's intention?") perspectives to distinguish between mental simulation (Theory-of-Mind reasoning) and Level-1 perspective-taking (Barresi & Moore, 1996).
- **Distractors with multiple difficulty levels**: Qtype 1 and 2 include three difficulty levels, with harder distractors being visually similar (e.g., comparable objects or spatial arrangements) and easier ones differing clearly in object type or environment. Qtype 4 uses three semantically distinct distractor types, ranging from low-level action descriptions to high-level intentions, to probe reasoning under varying cognitive demands.

3. Experiment

3.1. Setup

Inference: We evaluated 61 multimodal large language models (MLLMs) spanning diverse architectures, parameter scales (1B–110B), and training methodologies, including proprietary models (e.g., ChatGPT, Claude) and state-of-the-art open-source models (e.g., InternVL, Qwen, DeepSeek). Proprietary models were assessed via API calls, while open-source models were run locally using an 8×NVIDIA A100 80GB GPU cluster, with GPU allocation varying by model size. Official inference codebases were used to ensure reproducibility, and a unified evaluation toolkit was developed to handle multimodal inputs consistently (details in Appendix).

Evaluation: Model responses were compared to ground truth using template matching, with semantic matching by Llama-3.1-70B-Instruct applied when template matching failed. To mitigate answer-position bias, circular evaluation was employed, requiring correct answers across all permutations of multiple-choice options (details in Appendix).

3.2. Main Results

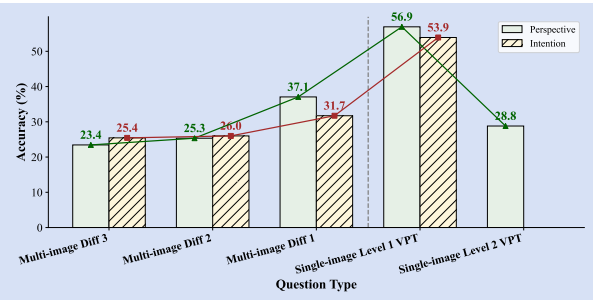

Figure 2: Comparative result between perspective taking and intention understanding across different difficulty levels and input types.

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 2. 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 2). 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 & 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.

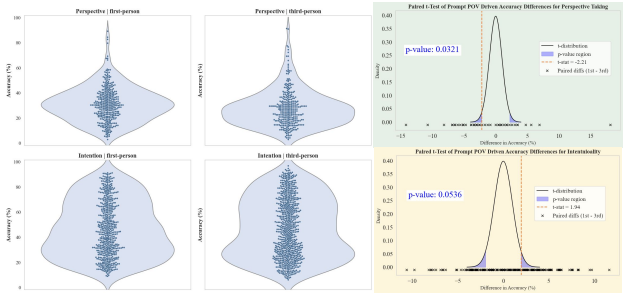

Figure 3: **Left:** Distribution of accuracy partitioned by probing concept and point-of-view of prompt; **Right:** Paired-T test results of single-image question for 2 types of prompts

3.3. 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.

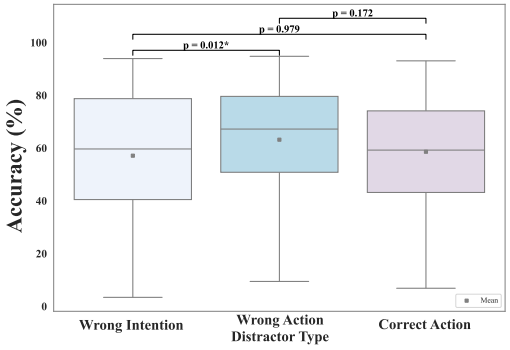

Figure 4: Accuracy by distractor type in Qtype 4 Ablation Test where the distractor type is controlled

Figure 4 reveals that average model accuracy varies across distractor types. Compared to the original Qtype 4 setup with an average accuracy of 53.9% (Figure 2), 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.

3.4. Benchmark Results

Benchmark results of representative models across Qtype subtasks are documented in Table 1 in Section 3.5

3.5. Benchmark Results

Differentiation effect is stronger on easier tasks. Top performers like `llava-video-72b-qwen2` (Zhang et al., 2024) achieve near-perfect accuracy (98% on Qtype 1 Difficulty 1), while weaker models remain below 30%. As task difficulty increases, accuracy variance decreases. Stronger models converge to similar low performance and weaker models consistently underperform across all levels.

Certain model series consistently excel, such as `qwen2_5_vl` series and `llava-video` series, of-

Table 1: Accuracy of selected models on each Qtype subtask. Best cells are bold and both best and second-best are shaded.

	Qtype 1 *Ego-Exo Match*			**Qtype 2** *Intention Match*			**Qtype 3** *Perspective Inference*	**Qtype 4** *Intention Inference*
Model	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%

ten scoring above 50% at larger scales (Team, 2025; Zhang et al., 2024). In contrast, the `eagle_series_x4` and `x5` models broadly underperform, with even the `13B` variant averaging below 20%, indicating a deficit in VPT knowledge from their architecture, pretraining, or fine-tuning (Shi et al., 2024).

Larger models generally perform better, as seen in the `vila` series (`vila1.5-40b` at 48% versus `vila1.5-3b` at 33%) (Lin et al., 2023). However, diminishing returns are evident in some cases, such as `llava-video-72b-qwen2`, which only slightly outperforms its `7B` counterpart (52% vs. 50%), suggesting that scaling beyond a certain point yields limited benefits in the hierarchy of perspective-taking abilities (Zhang et al., 2024).

4. 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 and some in and intention inference, they consistently struggle with Level-2 visual perspective-taking. 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 nor world model—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.

References

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

Barlassina, L. and Gordon, R. M. Folk psychology as mental simulation. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017. URL https://plato.stanford.edu/archives/sum2017/entries/folkpsych-simulation/.

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

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

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

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

Gallese, V. 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, 2007.

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

Gao, Q., Li, Y., Lyu, H., Sun, H., Luo, D., and Deng, H. Vision language models see what you want but not what you see, 2025. URL https://arxiv.org/abs/2410.00324.

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

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., Damen, D., Engel, J., Farinella, G. M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C., Newcombe, R., Park, H. S., Rehg, J. M., Sato, Y., Savva, M., Shi, J., Shou, M. Z., and Wray, M. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

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

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL https://arxiv.org/abs/1902.09506.

Jia, B., Chen, Y., Huang, S., Zhu, Y., and Zhu, S.-C. Lemma: A multiview dataset for learning multi-agent multi-view activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

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

Kessler, K. and Rutherford, H. The two forms of visuo-spatial perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in Psychology*, Volume 1 - 2010, 2010. ISSN 1664-1078. doi: 10.3389/fpsyg.2010.00213. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2010.00213.

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

Li, Y., Gao, Q., Zhao, T., Wang, B., Sun, H., Lyu, H., Luo, D., and Deng, H. Core knowledge deficits in multi-modal language models, 2025. URL https://arxiv.org/abs/2410.10855.

Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

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

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

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

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788, 2018. doi: 10.1109/CVPR.2018.00915.

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

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526, 1978. doi: 10.1017/S0140525X00076512.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., and Kanske, P. 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, 2021.

Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Huang, D.-A., Yin, H., Sapra, K., Yacoob, Y., Shi, H., Catanzaro, B., Tao, A., Kautz, J., Yu, Z., and Liu, G. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv:2408.15998*, 2024.

Team, Q. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

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

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

Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., and Li, C. Video instruction tuning with synthetic data, 2024. URL https://arxiv.org/abs/2410.02713.

Appendix

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. Table 2 lists all scenarios and tasks considered.

Table 2: 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 3). 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 4).

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 3: 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 4: 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 5, 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 5: 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 5: 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. **Manual Quality check**

   - 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 5). 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 6: 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. Experiment Details

B.1. Inference

We evaluated 61 models spanning both commercial closed-source systems and publicly available open-source models. Closed-source models were accessed via API and tested on local machines, whereas open-source models were downloaded from Hugging Face or GitHub and deployed on GPU servers for inference.

These models varied widely in architecture and size, ranging from 1B to 110B parameters (with the upper bound applying only to open-source models). Inference was conducted on computing clusters equipped with 8×NVIDIA A100 80 GB GPUs. Typically, models up to 13B could be run on a single GPU. Those between 13B and 32B used two GPUs, models between 32B and 70B used four GPUs, and the largest models required the full 8-GPU setup.

Models were categorized into three types based on supported input: single-image, multi-image, and video models. Each type was evaluated under corresponding experimental conditions. Video models underwent the most comprehensive tests, covering both video and single-image tasks. For multi-image settings, videos were decomposed into frame sequences fed as inputs. In the single-image condition, we restricted evaluation to tasks involving only a single image, excluding any requiring multi-frame or video data.

B.2. Evaluation

To assess model performance, we developed a robust answer-matching methodology that could accommodate diverse prompt formats and generative output variations. We tested several matching strategies and ultimately devised a hybrid method that balances template-based precision with semantic flexibility.

We examined four matching techniques:

1. **Exact Match**: This method cleans the output by removing special characters and performs a case-insensitive exact string match between model output and answer choices.
2. **"In" Match**: After cleaning, this method splits the model output into tokens and verifies if it contains exactly one valid answer choice.
3. **Template Match**: This approach matches outputs to specific answer patterns (e.g., "Answer: [choice]" or "[choice]. [explanation]"), requiring iterative adjustments to template coverage.
4. **LLM Match**: Using Llama3.1-70B and DeepSeek in an LLM-as-a-judge setup, this method provides the model with the full question prompt, choices, and visual/textual summaries, asking it to infer which choice the MLLM output aligns with.

We validated these approaches through 1) randomly sampling data points and examining their matching accuracy using each method, and 2) calculating the overall rate of each method.

Exact and "In" matchers suffered high fail rates, especially for complex outputs or models trained with explanation-oriented reasoning. Template matching, though more accurate for formatted responses, struggled with coverage and exceptions even after exhaustive template tuning. LLM match demonstrated strong performance in semantically mapping explanation-heavy outputs to choices, especially when the model used concessions or nuanced language—but it also risked hallucinating decisions, especially in short-output or noisy contexts.

To mitigate the weaknesses of each approach, we introduced a **Merge Match** mechanism. This strategy first attempts template matching and falls back to LLM match when no template match is found, thus combining the strengths of rule-based and semantic evaluation.

For all experiments, we adopted a zero-shot prompting strategy of the form $Q(M)T \rightarrow A$, where question text (Q), task description (T), and multiple-choice options (M) are provided as input, and the model generates the answer (A). Notably, to account for potential bias in multi-choice ordering, we applied *circular evaluation*, where the correct answer is systematically rotated through each choice slot. Only when the model consistently identifies the correct answer across all rotations is it marked correct, following the method proposed in Liu et al. (2023).