Seeing from Another Perspective: Evaluating Multi-View Understanding in MLLMs

Chun-Hsiao Yeh*[1], Chenyu Wang*[2,3,7], Shengbang Tong[4], Ta-Ying Cheng[5], Ruoyu Wang[3], Tianzhe Chu[2], Yuexiang Zhai[1], Yubei Chen[6], Shenghua Gao[2,3], Yi Ma[1,2]

[1]University of California, Berkeley
[2]The University of HongKong
[3]Transcengram
[4]New York University
[5]University of Oxford
[6]University of California, Davis
[7]SLAI

Abstract

Multi-view understanding, the ability to reconcile visual information across diverse viewpoints for effective navigation, manipulation, and 3D scene comprehension, is a fundamental challenge in Multi-Modal Large Language Models (MLLMs) to be used as embodied agents. While recent MLLMs have shown impressive advances in high-level reasoning and planning, they frequently fall short when confronted with multi-view geometric consistency and cross-view correspondence. To comprehensively evaluate the challenges of MLLMs in multi-view scene reasoning, we introduce *All-Angles Bench*, a human carefully benchmark with over **2,100** question-answer pairs from **90** diverse, real-world scenes. Our broad evaluation across **38** general-purpose and 3D spatial reasoning MLLMs reveals a substantial performance gap compared to humans. More critically, our analysis identifies two root failure modes: (1) *cross-view object mismatch*—the inability to establish consistent object correspondence across views; and (2) *cross-view spatial misalignment*—the failure to infer accurate camera poses and spatial layouts. These findings underscore a lack of multi-view awareness in current MLLMs, calling for architectural innovations beyond prompt tuning alone. We believe that our benchmark offers valuable insights toward building spatially-intelligent MLLMs.

Introduction

Multi-view understanding is a fundamental challenge in bridging machine and human-level understanding (Das et al. 2018; Yu et al. 2019; Hong et al. 2023) because it underpins an agent's ability to perceive the environment consistently from diverse viewpoints. By ensuring geometric coherence and cross-view consistency, agents can accurately reconstruct scene layouts and object relationships — capabilities critical for effective navigation, manipulation, and interaction in the real world (Song et al. 2022; Suglia et al. 2021). The recent advancement in Multimodal Large Language Models (MLLMs) demonstrates strong capabilities in high-level reasoning and task planning (Li et al. 2024b; Hurst et al. 2024; Team et al. 2023; Bai et al. 2025; Chen et al. 2024d), and thus the feasibility of directly using MLLMs as embodied agents is an intriguing research challenge (Huang et al. 2022; Yue et al. 2024a; Kim et al. 2024; Liu et al. 2024b). However, such capacities alone are insufficient for generalist embodied agents operating in the real world, where a comprehensive 3D scene understanding and robust multi-view reasoning are pivotal (Jia et al. 2024; Cheng et al. 2025). Recent studies survey that MLLMs lacking multi-view scene understanding often commit agent manipulation and navigation errors such as misjudge the target distance, skip partially occluded obstacles, stemming from limited awareness of multi-view geometry and object relationships (Yu et al. 2025; Zhu et al. 2024a). Since these models must navigate, manipulate, and make decisions in real world environments, it is vital to evaluate (and ultimately strengthen) their multi-view understanding capabilities. Yet, this aspect remains underexplored in details.

To this end, we raise two questions: *(1) Do MLLMs possess the ability to understand multiple viewpoints simultaneously?* and *(2) What are the key challenges in MLLMs to gain better multi-view understanding?*

To address these questions and in light of the lack of benchmarks to evaluate multi-view reasoning, we introduce *All-Angles Bench*, comprising over 2,100 carefully human-annotated question-answer pairs across 90 diverse multi-view scenes in real world (Grauman et al. 2024; Khirodkar et al. 2023). We define six tasks — *counting, attribute identification, relative distance, relative direction, manipulation, and camera pose estimation* — with a focus on evaluating MLLM's geometric understanding and its ability to align information consistently across multi-view scenes. To better evaluate whether models truly possess multi-view capabilities, we also propose a paired question scheme by creating a second question with the same content but with slightly changed wording/order of views. We benchmark 38 representative MLLMs (including Gemini-2.5-Flash, Claude-4 Sonnet, and GPT-4.1) against human evaluators. As revealed in Figure 1, a substantial performance gap persists between current MLLMs and human evaluators.

To better understand why MLLMs fall short of human-level multi-view reasoning, we conduct an in-depth analy-

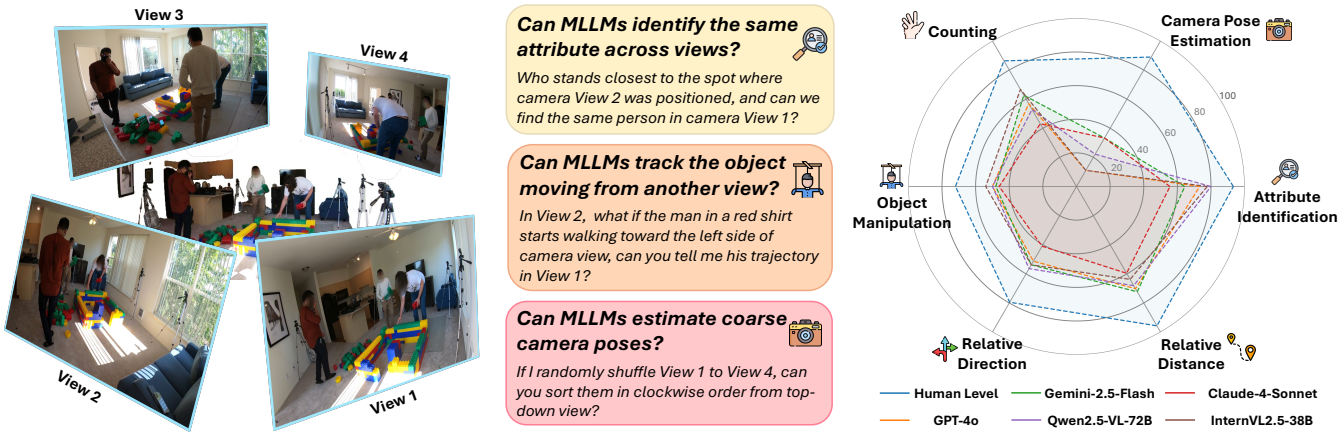

Figure 1: We present *All-Angles Bench*, a human-annotated benchmark with over 2,100 Q&A pairs from 90 diverse scenes for evaluating multi-view understanding of MLLMs. **Left and Middle:** An example question setup of multiple views capturing the same scene and the corresponding questions. **Right:** Accuracies of six notable MLLMs across different question categories.

sis of commonly failed questions and tasks, and derive two key findings. *First, MLLMs struggle to identify the same object across multiple views.* We further test whether chain-of-thought prompting, a technique that has proven effective in other reasoning tasks (Yang et al. 2024a; Zhang et al. 2025; Rudman et al. 2025), could address this limitation. However, our experiments reveal that these linguistic strategies do not provide consistent improvements across models for multi-view reasoning. This suggests that more fundamental domain-specific refinements to multi-view awareness modules or training data are necessary for MLLMs to fully internalize cross-view consistency. *Second, MLLMs often fail to establish correspondence between different viewpoints.* We visualize how models infer scene layouts from multiple perspectives, revealing a consistent inability to accurately estimate camera poses, which in turn impedes performance on tasks like *relative direction* and *object manipulation*. We hope these insights will be helpful to future research towards bringing more better multi-view capabilities in MLLMs.

All-Angles Bench

Overview of All-Angles Bench

Most existing benchmarks to evaluate MLLMs primarily rely on single-view or egocentric data, leaving the critical capabilities of multi-view consistency and correspondence largely unexamined. To address this gap, we introduce *All-Angles Bench*, a comprehensive platform designed to rigorously assess the geometric understanding of MLLMs. Sourced from real-world datasets Ego-Exo4D (Grauman et al. 2024) and EgoHumans (Khirodkar et al. 2023), our benchmark consists of 2,132 multiple-choice question-answer pairs across 90 diverse scenes. It specifically probes a model's ability to reason about geometry and align information across perspectives through six targeted task categories: ***(1) Counting***, ***(2) Attribute Identification***, ***(3) Relative Distance***, ***(4) Relative Direction***, ***(5) Object Manipulation***, and ***(6) Camera Pose Estimation***. The scenes encompass a wide range of activities (e.g., basketball, soccer, cooking, music playing) and environments (e.g., offices, gym, repair store, kitchen, playground) to ensure broad real-world relevance. Furthermore, we generate paired questions that alter viewpoints while preserving semantic correspondence, explicitly testing for robust reasoning over superficial cues. As shown in Figure 2, each question is designed to provide a challenging yet realistic test for evaluating MLLMs' geometric understanding and multi-view correspondence.

Benchmark Collection Process

We build a benchmark collection pipeline to effectively generate high quality question-answer pairs for multi-view understanding, as shown in Figure 3. To ensure the benchmark quality, all questions were manually annotated by human annotators after collecting and clipping the raw questions.

Data Collection & Question Type Design. Our benchmark builds on 90 diverse multi-view scenes from Ego-Exo4D and EgoHumans. Spanning a broad spectrum of real-world activities and environments, each scene includes footage from at least three viewpoints to ensure a rich context for multi-view analysis. Upon this data, we manually designed six task categories that target fundamental aspects of spatial intelligence. These tasks evaluate a range of skills, from object correspondence (*counting, attribute identification*) and spatial relationships (*relative distance, relative direction*) to geometric and dynamic reasoning (*object manipulation, camera pose estimation*). Appendix details the specific design.

Question Creation & Human Annotation. Our question generation follows a human-in-the-loop pipeline. We first leverage GPT-4o to produce initial questions grounded in the multi-view scenes. This draft then undergoes meticulous human validation and refinement. Annotators correct logical inconsistencies (e.g., contradictory object descriptions), resolve ambiguous answer choices, and annotate the single ground-truth answer (Figure 3, middle). For example, in attribute identification, the MLLM might inconsistently describe an object across two different camera views. In rel-

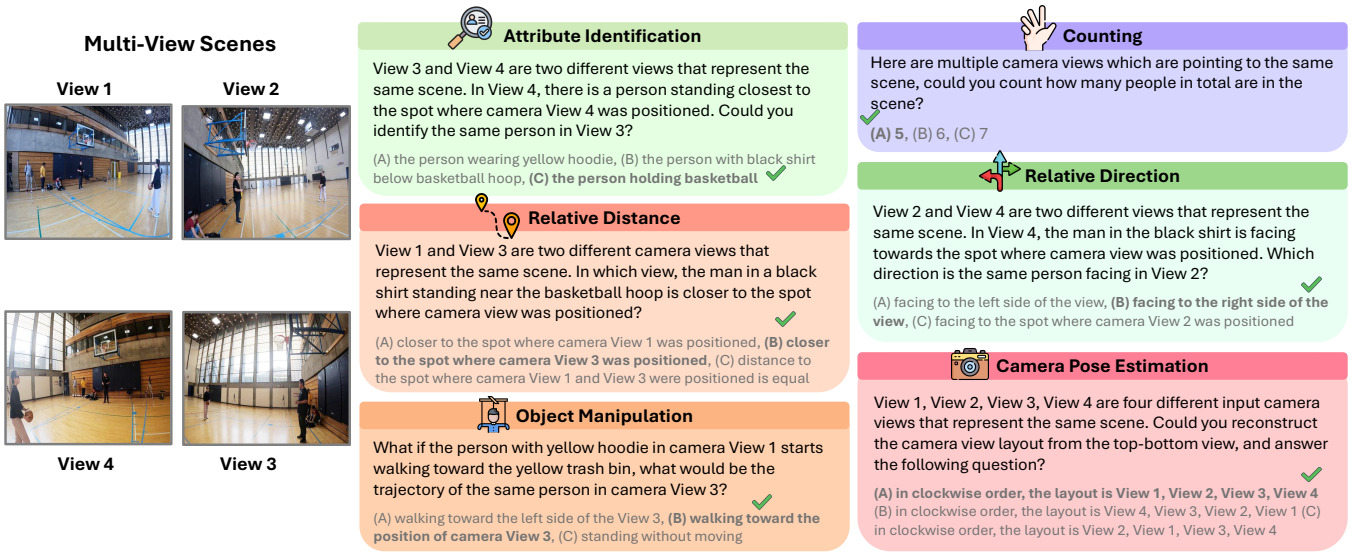

Figure 2: Overview of *All-Angles Bench*. Our benchmark targets a comprehensive view of multi-view understanding, spanning six primary question types. These question types are designed to investigate several major aspects of 3D scene understanding, from creating correspondence between objects to associating relative object and camera poses.

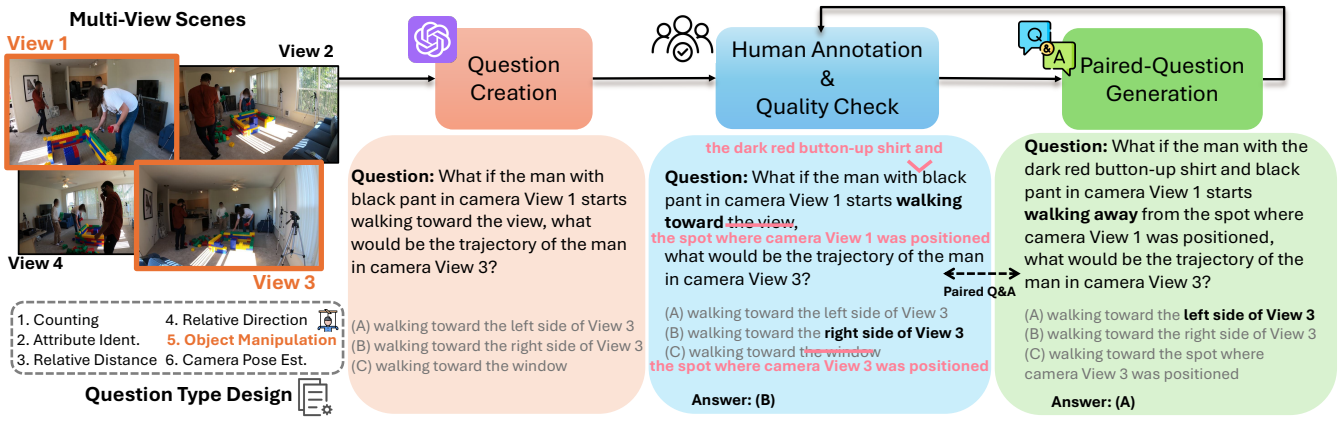

Figure 3: The construction pipeline for *All-Angles Bench*. (1) We curate 90 diverse multi-view scenes and design six tasks focused on geometric reasoning and cross-view consistency. (2) An MLLM generates initial questions, which then undergo rigorous human annotation to ensure correctness and clarity. (3) To test model robustness, we systematically create paired questions by altering view perspectives while preserving semantic meaning, followed by a final quality check.

ative direction, the MLLM might inconsistently describe an object or offer contradictory options (e.g., *"facing the right side"* vs. *"facing the stove"*). This human-led stage, detailed in the Appendix, is critical for the benchmark's clarity and correctness. We tailor inputs for each task; *counting* and *camera pose estimation* use all views, while other tasks use two randomly selected viewpoints.

Paired-Question Generation & Human Quality Check. To rigorously test for robust understanding beyond simple accuracy, we generate paired questions. This involves systematically altering a query's perspective—for instance, by swapping view references (e.g., View 1 vs. View 2) or inverting directional language—while preserving the core visual correspondence. This process, inspired by language manipulation techniques (Yang et al. 2024b; Zhu et al. 2024b), requires careful verification of view-to-view geometric consistency. Therefore, a final human quality check is performed to ensure the logical and geometric alignment of each pair. This results in 85.3% of relevant questions having a validated counterpart (the *counting* task is not involved). This pairing strategy, with final statistics shown in Figure 4, allows us to test whether MLLMs genuinely understand multi-view scenarios or merely guess answers.

MLLMs Have Multi-View Understanding?

Benchmark Models and Human Evaluation. We evaluate a broad spectrum of MLLMs, including closed-source mod-

Methods	Avg.	Attribute	Cam. Pose	Counting	Manipul.	Rel. Dir.	Rel. Dist.
		Multiple-Choice Answer					
Perf. Against Human (250 Q&As)							
Human Level	82.0	93.3	88.9	86.3	72.0	79.5	95.7
GPT-4o	56.8	73.3	11.1	56.9	50.0	51.3	70.2
Gemini-2.5-Flash	58.4	64.4	33.3	62.7	48.0	53.8	72.3
Claude-4-Sonnet	48.0	55.6	33.3	43.1	46.0	41.0	59.6
InternVL2.5-38B	59.6	77.8	11.1	66.7	54.0	53.8	63.8
Qwen2.5-VL-72B	58.4	80.0	22.2	52.9	50.0	56.4	68.1
Closed-source Models							
GPT-4o	51.4	71.0	27.3	55.4	41.4	40.9	59.9
GPT-4.1	53.6	76.2	38.1	55.4	48.7	36.6	57.3
Gemini-1.5-Pro	43.6	61.6	25.0	39.4	40.3	29.8	51.4
Gemini-2.0-Flash	51.8	65.8	28.4	66.5	45.0	38.1	58.1
Gemini-2.5-Flash	57.3	73.6	34.1	58.2	48.5	49.4	66.4
Claude-3.5-Sonnet	49.7	68.7	27.8	46.2	41.8	41.8	57.7
Claude-3.7-Sonnet	49.3	70.0	37.5	42.2	39.9	45.7	52.8
Claude-4-Sonnet	48.9	65.5	29.5	38.2	47.7	36.1	58.7
Open-source Models							
DeepSeek-VL2-Small	46.2	66.8	36.9	36.7	44.8	32.4	49.6
DeepSeek-VL2	45.7	71.5	25.6	36.7	47.7	30.1	46.8
InternVL2.5-2B	44.1	59.3	38.1	39.0	46.6	25.9	47.8
InternVL2.5-4B	47.1	66.6	36.9	47.0	40.1	33.2	52.2
InternVL2.5-8B	49.4	73.6	31.8	51.4	43.7	34.1	52.2
InternVL2.5-38B	53.1	79.4	31.3	56.2	46.2	42.6	53.2
InternVL2.5-78B	52.4	82.8	38.6	56.2	42.2	38.6	51.6
InternVL3-2B	48.5	66.6	42.0	47.0	43.3	38.1	49.8
InternVL3-8B	50.7	78.6	36.4	50.9	42.6	34.6	53.2
InternVL3-38B	57.6	82.2	36.9	50.2	48.3	52.6	61.9
Qwen2.5-VL-3B	43.1	63.2	26.1	44.6	34.7	35.5	46.4
Qwen2.5-VL-32B	54.0	78.3	32.4	54.2	49.8	40.9	56.3
Qwen2.5-VL-72B	54.8	81.7	34.1	55.4	45.0	48.3	55.3
Ovis2-2B	46.3	62.4	48.3	50.2	43.9	28.4	46.0
Ovis2-4B	48.1	63.5	27.3	55.8	43.5	36.4	52.4
Ovis2-8B	48.5	73.4	27.3	51.4	43.1	33.8	50.8
Ovis2-16B	51.2	76.0	23.9	53.4	47.5	40.1	52.0
Ovis2-34B	55.2	80.2	29.6	53.4	49.0	43.8	59.9
Cambrian-8B	29.8	51.7	8.5	34.7	30.7	30.4	16.8
Cambrian-13B	39.3	56.1	23.9	30.7	36.8	31.5	43.9
Cambrian-34B	42.5	61.1	26.7	37.9	39.5	36.9	42.7
LLaVA-Onevision-Qwen2-7B	45.2	64.2	20.5	41.0	42.4	36.4	50.2
LLaVA-Onevision-Qwen2-72B	50.2	78.3	20.5	42.2	47.7	33.8	57.3
LLaVA-Video-Qwen2-7B	42.9	65.5	14.8	37.1	42.9	31.0	47.0
LLaVA-Video-Qwen2-72B	50.2	75.7	28.4	40.6	45.4	42.3	53.2

Table 1: Evaluation results for 33 MLLMs. We consolidate performance from both closed-source and open-source MLLMs. We use deeper-gray and light-gray to highlight the best and second-best result among all models in each sub-task.

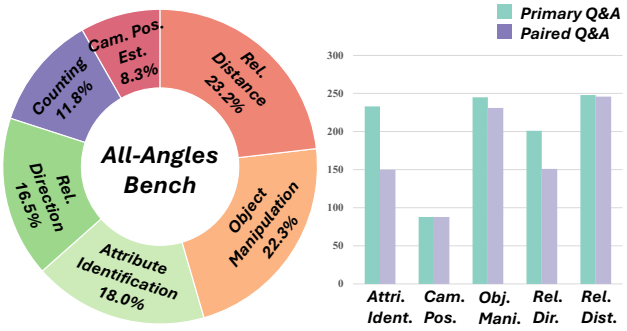

Figure 4: Statistical overview of *All-Angles Bench*. The pie chart shows the distribution of 6 sub-tasks of multi-view understanding. The bar plot illustrates the breakdown by primary and paired question-answers of each sub-task.

	Avg.	Attr.	C. Pos	Count	Mani.	Dir.	Dist.
3D Spatial Reasoning Models							
VG-LLM (Zheng et al. 2025)	33.7	56.7	16.5	26.3	30.0	26.9	34.2
AoTD (Shi et al. 2025)	36.8	41.5	32.4	27.9	37.6	26.7	45.5
VLM-3R (Fan et al. 2025)	40.3	56.1	22.7	39.4	35.9	30.9	45.6
CoF (Ghazanfari et al. 2025)	47.8	75.7	35.8	41.4	38.7	38.7	49.0
SpaceR (Ouyang et al. 2025)	49.7	72.8	51.1	46.2	41.2	38.1	49.4

Table 2: 3D Spatial Reasoning MLLMs. Evaluation of recent 3D spatial reasoning MLLMs on *All-Angles Bench* across 6 spatial sub-tasks.

els like Gemini, Claude, and GPT, and open-source models like Deepseek-VL2 (Wu et al. 2024), Qwen2.5-VL (Bai et al. 2025), InternVL (Chen et al. 2024c; Zhu et al. 2025), Cambrian (Tong et al. 2024a), LLaVA-OneVision (Li et al. 2024c), LLaVA-NeXT-Video (Zhang et al. 2024), Ovis2 (Lu et al. 2024), and some 3D spatial reasoning MLLMs. We use standard protocols with temperature set to zero. We also perform human evaluation on a 250-question subset from our *All-Angles Bench*, covering all six task categories, with each question answered independently by annotators. For fair comparison, we report performance for Gemini-2.5-

Flash, Claude-4-Sonnet, GPT-4o, Qwen2.5-VL-72B, and InternVL2.5-38B on this subset.

Results on General and Spatial Reasoning MLLMs & Findings. The results in Table 1 and Table 2 reveal a stark reality: a substantial performance chasm separates all evaluated MLLMs from human-level multi-view understanding. We highlight two key findings that characterize this gap.

> ***Finding 1:*** *Simple task for human like coarse camera pose estimation poses challenges for MLLMs.*

While humans achieve near-perfect accuracy on several of our benchmark's tasks, MLLMs struggle profoundly with fundamental geometric inference. This is most evident in camera pose estimation, where humans score 88.9%. In sharp contrast, top-tier models like Gemini-2.5-Flash and Qwen2.5-VL-72B lag by **over 50% margins** as well as many open-source models. This demonstrates a core failure in the ability of current MLLMs to reconcile geometric information across different viewpoints.

> ***Finding 2:*** *Specialized open-source models show pockets of excellence in orientation-sensitive tasks.*

Interestingly, certain open-source models outperform leading closed-source competitors on tasks requiring orientation and trajectory reasoning. On *object manipulation*, Qwen2.5-VL-32B (49.8) surpasses Gemini-2.5-Flash (48.5) and Claude-4-Sonnet (47.7). We observe that Qwen2.5-VL-72B integrates robust video understanding and fine-grained visual grounding modules (as highlighted in its model report). Similarly, on *attribute identification*, InternVL2.5-78B (82.8) leads all other models. We hypothesize this stems from their specialized, video-focused training regimes that emphasize fine-grained visual grounding and orientation tracking. This suggests that for complex spatial reasoning, targeted architectural and training refinements may be more crucial than model scale alone.

> ***Finding 3:*** *3D specialized spatial reasoning MLLMs close the gap — but often only limited to the subtasks.*

Recent MLLMs purpose-built for spatial reasoning—like SpaceR (Ouyang et al. 2025), VLM-3R (Fan et al. 2025), and AoTD (Shi et al. 2025)—achieve gains across spatial subtasks. SpaceR, in particular, scores 51.1 on camera pose estimation and surpasses many general-purpose MLLMs. However, their strength is often limited to the subtasks they explicitly target. While promising, these results suggest that simply injecting spatial priors or specialized architectures helps but does not solve the general multi-view challenge.

Robustness on Paired Questions. Single-question accuracy fails to capture true reasoning versus pattern-matching. We probe this by evaluating consistency on paired questions—semantically equivalent queries from different viewpoints or with altered phrasing—to test if a model's understanding is stable. We classify outcomes into three categories: 1) ***CC (Both Correct)***; 2) ***WW (Both Wrong)***; and 3) ***IC (Inconsistent)*** if it passes one but fails the other. A high ***IC*** rate suggests brittle success, not genuine multi-view understanding, as simple perspective changes lead to failure.

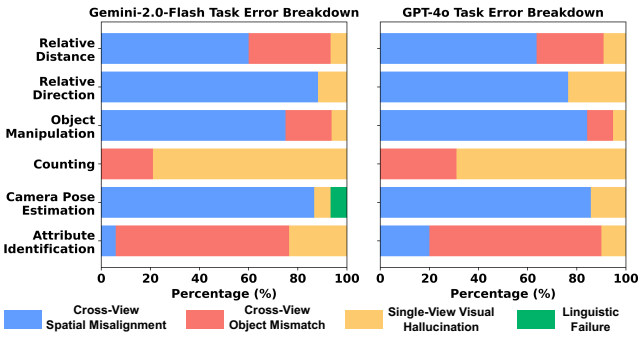

Figure 5: MLLM error breakdown. For both Gemini-2.0-Flash and GPT-4o, the majority of failures fall into cross-view spatial misalignment and object mismatch.

As shown in Figure 6, we report ***IC*** (inconsistent) outcomes across six leading MLLMs—three open-source (Ovis2-34B, Qwen2.5-VL-72B, InternVL2.5-38B) and three closed-source (GPT-4o, Gemini-2.0-Flash, Claude-3.7-Sonnet). Observations: 1) GPT-4o exhibits high ***IC*** (around 70%) on relative distance tasks, while others are around 40%. 2) All models struggle with relative direction, with ***IC*** rates surpassing 40%, highlighting the challenge of multi-view orientation. 3) Gemini-2.0-Flash and Claude-3.7-Sonnet show balanced inconsistency, while Ovis2-34B and GPT-4o vary significantly by task.

Why Do MLLMs Struggle with Multi-View Understanding?

To investigate specific weaknesses of MLLMs in multi-view comprehension, we evaluate each question type in our *All-Angles Bench*. We select the top-performing closed-source and open-source MLLMs in our benchmark and systematically identify where these models succeed or fail in understanding multi-view scenarios.

To provide further statistical support, we conducted a detailed human-annotated error analysis on QA failures from Gemini-2.0-Flash and GPT-4o for all tasks in Figure 5. Each failure was assigned to one of four mutually exclusive error categories: *1) Cross-View Spatial Misalignment*, *2) Cross-View Object Mismatch*, *3) Single-View Visual Hallucination*, and *4) Linguistic Failure*, with annotation based on the earliest and most dominant error observed. Figure 5 shows that the majority of failures fall into ***Cross-View Object Mismatch*** and ***Cross-View Spatial Misalignment***, demonstrating MLLMs fail in maintaining geometric consistency and object identity across viewpoints. These findings echo the conclusions of VSI-Bench (Yang et al. 2024a), which similarly identified spatial reasoning as a key limitation.

1) Cross-View Object Mismatch: The Failure of Multi-View Correspondence. We first investigate the multi-view counting task since we are curious about the discrepancy be-

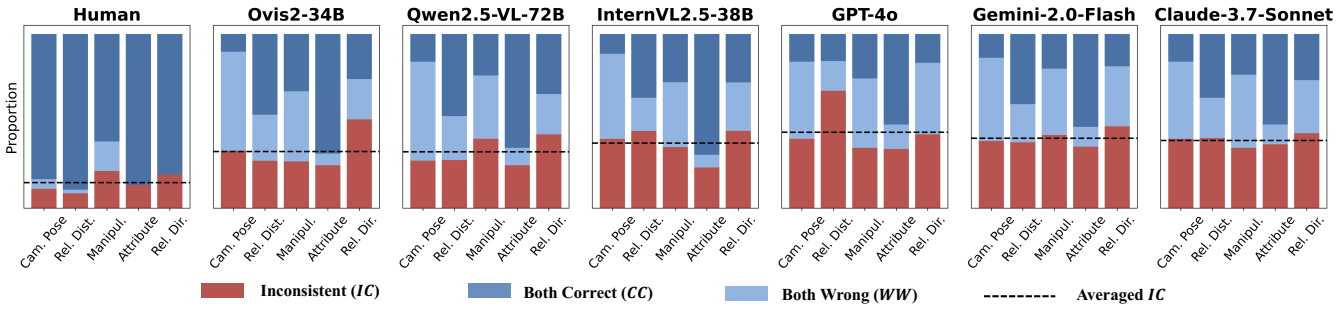

Figure 6: Paired question-answers inconsistency across 6 MLLMs. We report the proportions of ***IC***, ***CC***, and ***WW***. Notably, GPT-4o struggles with relative distance (around 70% inconsistency). Human, Gemini-2.0-Flash and Claude-3.7-Sonnet exhibit more balanced performance, whereas Ovis2-34B and GPT-4o vary considerably across tasks.

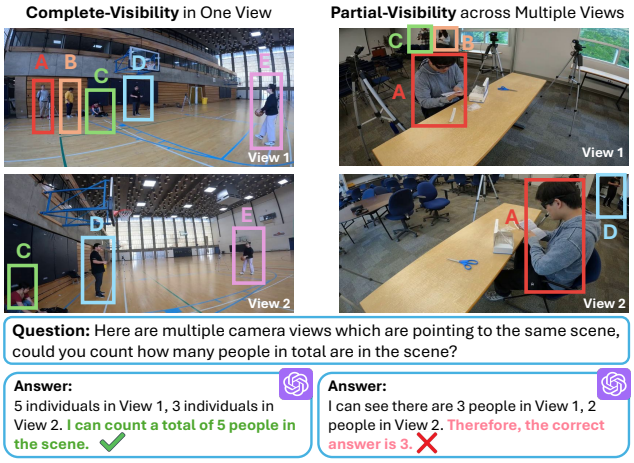

Figure 7: Complete- and Partial-visibility counting remains challenging for MLLMs, which often succeed with full visibility but may miscount by favoring the highest per-view number rather than integrating across views.

tween egocentric view and multi-view counting. While models perform well in ***complete-visibility in one view*** scenarios where all objects appear in at least one view, they consistently fail with ***partial-visibility across multiple views*** when they must reconcile entities across multi-views (e.g., Person A and B in View 1, and Person C and D in View 2). This failure often stems from flawed heuristics; for instance, as shown in Figure 7, GPT-4o may occasionally report the maximum count from any single view, a shortcut that bypasses the essential task of cross-view entity reconciliation.

Can Reasoning Injection Improve MLLM's Ability? We evaluate four prompting strategies—*Zero-Shot CoT*, *Self-Consistency with CoT*, our proposed *Identification CoT*, and *Coarse Correspondence* (Liu et al. 2024a), which leverages visual markers to image—on three representative MLLMs: GPT-4o, InternVL2.5-38B, and Ovis2-34B. *Identification CoT* guides the model to (1) describe each visible person, (2) align identities across views, and (3) count uniquely.

As shown in Table 3, GPT-4o benefits significantly

	View Type	Bl.	ZS-CoT	Self-Consist.	Ident.-CoT	Co. Corr.
GPT-4o	Compl. Vis.	65.5	63.6 (-1.9)	61.8 (-3.7)	69.1 (+3.6)	**80.0** (+14.5)
	Partial Vis.	41.8	50.9 (+9.1)	52.7 (+10.9)	61.8 (+20.0)	**67.3** (+25.5)
InternVL	Compl. Vis.	73.2	63.6 (-9.6)	61.8 (-11.4)	67.2 (-6.0)	**78.2** (+5.0)
	Partial Vis.	65.5	60.0 (-5.5)	61.8 (-3.7)	67.2 (+1.7)	**74.5** (+9.0)
Ovis	Compl. Vis.	65.5	61.8 (-3.7)	63.6 (-1.9)	67.2 (+1.7)	**76.4** (+10.9)
	Partial Vis.	60.0	54.5 (-5.5)	52.7 (-7.3)	63.6 (+3.6)	**74.5** (+14.5)

Table 3: Reasoning prompt analysis. CoT aids GPT-4o in partial-visibility cases but helps robust models (e.g., InternVL) less. These results indicate that prompt refinement is insufficient and visual markers may be required.

under partial visibility, with *Identification CoT* improving accuracy by +20% and *Coarse Correspondence* by +25.5%, highlighting that reasoning and visual grounding can compensate for occlusion. However, stronger models like InternVL2.5-38B gain little or even degrade with prompting—mirroring prior findings (Yang et al. 2024a) that CoT is less effective for spatially capable models such as Gemini-1.5. These results suggest that while prompt reasoning helps weaker MLLMs, its benefit plateaus with stronger ones. Moreover, approaches like *Coarse Correspondence* require additional effort to annotate visual markers before evaluation. Ultimately, advancing multi-view understanding likely demands architectural or training-level innovations, rather than prompt refinement alone.

2) Cross-View Spatial Misalignment: Failure with Camera Pose and Geometry. We observe that MLLMs struggle with *orientation-sensitive* challenges like camera pose and trajectory estimation (Table 1). To investigate this, we tasked GPT-4o and Gemini-2.0-Flash with inferring spatial layouts using a grid-based visualization prompt inspired by (Yang et al. 2024a).

As illustrated in Figure 8 (*object manipulation*) and Figure 9 (*camera pose estimation*), many orientation-related errors stem from the model's inability to reconcile viewpoint transformations. models frequently misalign camera coordinates or overlook background cues. This impairs down-

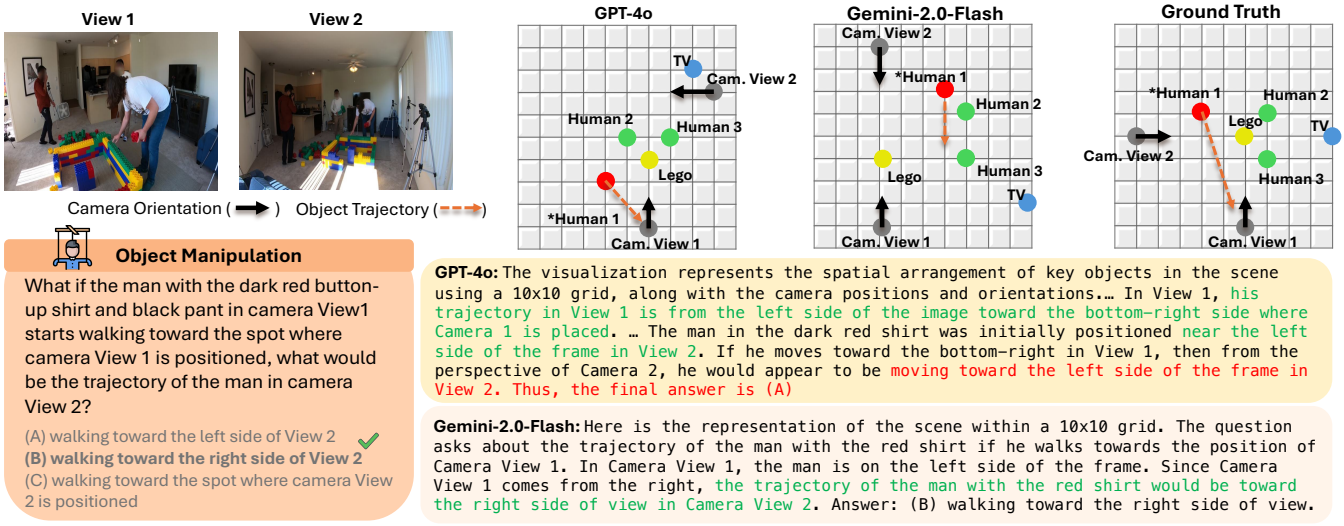

Figure 8: Visualization of multi-view scene reconstruction. While GPT-4o and Gemini-2.0-Flash show proficiency in single-view reconstruction, they struggle to align distinct camera perspectives. This misalignment leads to incorrect directional reasoning (e.g., trajectory tracking across views), revealing significant gaps in multi-view consistency.

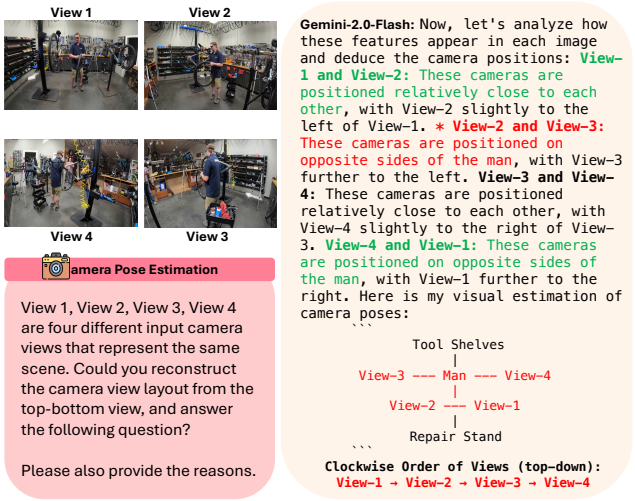

Figure 9: Visualization of camera pose estimation. When asked to order the camera poses in clockwise order, MLLMs fail completely despite providing detailed reasonings.

stream tasks requiring fine-grained geometric reasoning, such as *relative direction* or *object manipulation*. These observations echo our findings that domain-specific or architectural refinements, especially emphasizing viewpoint consistency may close the gap in multi-view understanding.

Related Works

Multimodal Large Language Models. MLLMs (Liu et al. 2023; Tong et al. 2024a; Bai et al. 2023; Hurst et al. 2024) show impressive capabilities (Yue et al. 2024b; Liu et al. 2024c; Tong et al. 2024c; Yue et al. 2024a; Li et al. 2024a; Zhai et al. 2024; Zhou et al. 2024; Tong et al. 2024b; Xu et al. 2024; Wang et al. 2024). An increasing number of studies (Hong et al. 2023; Chen et al. 2024a,b) are focusing on video understanding and embodied real-world tasks. Our work contributes to this area by: 1) providing a timely benchmark to assess multi-view perception—a fundamental capability for 3D and 4D tasks; and 2) analyzing why current models struggle with multi-view understanding.

Benchmarking Visual Spatial Ability. Recent works (Fu et al. 2024; Yang et al. 2024a; Li et al. 2024d) study the visual spatial ability of MLLMs. Our work is most relevant to VideoMME (Fu et al. 2024), VSI-Bench (Yang et al. 2024a), and MV-Bench (Li et al. 2024d), which primarily emphasize temporal reasoning (Fu et al. 2024; Li et al. 2024d) or egocentric spatial intelligence (Yang et al. 2024a). Our work focuses on multi-view understanding, a cornerstone for 3D and 4D reasoning. Unlike previous work assessing single-view or temporal reasoning, we explicitly evaluate how models align geometric and semantic information across multiple viewpoints. We further provide a breakdown analysis that dissects model deficiencies in multi-view understanding.

Conclusion

We introduce *All-Angles Bench* to evaluate MLLMs' multi-view understanding. Benchmarking 38 models across over 2,100 samples reveals critical limitations in geometric consistency and cross-view correspondence. These findings underscore the necessity of domain-specific training to bridge the gap toward human-level multi-view reasoning.

Acknowledgements

This work is supported by General Research Fund of the Research Grants Council (grant #17200725), the NSFC #62172279, and in part by the JC STEM Lab funded by The Hong Kong Jockey Club Charities Trust.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

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

Chen, Y.; Xue, F.; Li, D.; Hu, Q.; Zhu, L.; Li, X.; Fang, Y.; Tang, H.; Yang, S.; Liu, Z.; et al. 2024b. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2025. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37: 135062–135093.

Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; and Batra, D. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–10.

Fan, Z.; Zhang, J.; Li, R.; Zhang, J.; Chen, R.; Hu, H.; Wang, K.; Qu, H.; Wang, D.; Yan, Z.; et al. 2025. VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction. *arXiv preprint arXiv:2505.20279*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Ghazanfari, S.; Croce, F.; Flammarion, N.; Krishnamurthy, P.; Khorrami, F.; and Garg, S. 2025. Chain-of-Frames: Advancing Video Understanding in Multimodal LLMs via Frame-Aware Reasoning. *arXiv preprint arxiv:2506.00318*.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. 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*, 19383–19400.

Hong, Y.; Lin, C.; Du, Y.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2023. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9202–9212.

Huang, W.; Abbeel, P.; Pathak, D.; and Mordatch, I. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, 9118–9147. PMLR.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Jia, B.; Chen, Y.; Yu, H.; Wang, Y.; Niu, X.; Liu, T.; Li, Q.; and Huang, S. 2024. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, 289–310. Springer.

Khirodkar, R.; Bansal, A.; Ma, L.; Newcombe, R.; Vo, M.; and Kitani, K. 2023. Ego-humans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19807–19819.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

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

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024c. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024d. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Liu, B.; Dong, Y.; Wang, Y.; Rao, Y.; Tang, Y.; Ma, W.-C.; and Krishna, R. 2024a. Coarse Correspondences Elicit 3D Spacetime Understanding in Multimodal Language Model. *arXiv preprint arXiv:2408.00754*.

Liu, F.; Fang, K.; Abbeel, P.; and Levine, S. 2024b. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

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 *ECCV*.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Ouyang, K.; Liu, Y.; Wu, H.; Liu, Y.; Zhou, H.; Zhou, J.; Meng, F.; and Sun, X. 2025. SpaceR: Reinforcing MLLMs in Video Spatial Reasoning. *arXiv preprint arXiv:2504.01805*.

Rudman, W.; Golovanesky, M.; Bar, A.; Palit, V.; LeCun, Y.; Eickhoff, C.; and Singh, R. 2025. Forgotten Polygons: Multimodal Large Language Models are Shape-Blind. *arXiv preprint arXiv:2502.15969*.

Shi, Y.; Di, S.; Chen, Q.; and Xie, W. 2025. Enhancing Video-LLM Reasoning via Agent-of-Thoughts Distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8523–8533.

Song, C. H.; Kil, J.; Pan, T.-Y.; Sadler, B. M.; Chao, W.-L.; and Su, Y. 2022. One step at a time: Long-horizon vision-and-language navigation with milestones. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15482–15491.

Suglia, A.; Gao, Q.; Thomason, J.; Thattai, G.; and Sukhatme, G. 2021. Embodied bert: A transformer model for embodied, language-guided visual task completion. *arXiv preprint arXiv:2108.04927*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S. C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; et al. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*.

Tong, S.; Fan, D.; Zhu, J.; Xiong, Y.; Chen, X.; Sinha, K.; Rabbat, M.; LeCun, Y.; Xie, S.; and Liu, Z. 2024b. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024c. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*.

Wang, C.; Luo, W.; Chen, Q.; Mai, H.; Guo, J.; Dong, S.; Li, Z.; Ma, L.; Gao, S.; et al. 2024. Mllm-tool: A multimodal large language model for tool agent learning. *arXiv preprint arXiv:2401.10727*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xu, J.; Zhao, Z.; Wang, C.; Liu, W.; Ma, Y.; and Gao, S. 2024. Cad-mllm: Unifying multimodality-conditioned cad generation with mllm. *arXiv preprint arXiv:2411.04954*.

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

Yang, Y.; Zhang, S.; Shao, W.; Zhang, K.; Bin, Y.; Wang, Y.; and Luo, P. 2024b. Dynamic Multimodal Evaluation with Flexible Complexity by Vision-Language Bootstrapping. *arXiv preprint arXiv:2410.08695*.

Yu, H.; Li, W.; Wang, S.; Chen, J.; and Zhu, J. 2025. Inst3D-LMM: Instance-Aware 3D Scene Understanding with Multi-modal Instruction Tuning. arXiv:2503.00513.

Yu, L.; Chen, X.; Gkioxari, G.; Bansal, M.; Berg, T. L.; and Batra, D. 2019. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6309–6318.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024b. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*.

Zhai, Y.; Bai, H.; Lin, Z.; Pan, J.; Tong, S.; Zhou, Y.; Suhr, A.; Xie, S.; LeCun, Y.; Ma, Y.; et al. 2024. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. In *NeurIPS*.

Zhang, J.; Yao, D.; Pi, R.; Liang, P. P.; et al. 2025. VLM 2-Bench: A Closer Look at How Well VLMs Implicitly Link Explicit Matching Visual Cues. *arXiv preprint arXiv:2502.12084*.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zheng, D.; Huang, S.; Li, Y.; and Wang, L. 2025. Learning from Videos for 3D World: Enhancing MLLMs with 3D Vision Geometry Priors. *arXiv preprint arXiv:2505.24625*.

Zhou, C.; Yu, L.; Babu, A.; Tirumala, K.; Yasunaga, M.; Shamis, L.; Kahn, J.; Ma, X.; Zettlemoyer, L.; and Levy, O. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Zhu, C.; Wang, T.; Zhang, W.; Pang, J.; and Liu, X. 2024a. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhu, K.; Wang, J.; Zhao, Q.; Xu, R.; and Xie, X. 2024b. Dynamic evaluation of large language models by meta probing agents. *arXiv preprint arXiv:2402.14865*.