OmniCoT: A Benchmark for Global and Multi-Step Panoramic Reasoning

Haocong He[1,*], Chenfei Liao[2,*,†], Zichen Wen[1,7], Zihao Dongfang[2], Xu Zheng[2], Bin Ren[3], Chang Su[4], Zixin Zhang[2], Harold H. Chen[2], Hongfei Zhang[2], Weijia Li[5,7], Kailun Yang[6], Conghui He[7], Xuming Hu[2,9], Nicu Sebe[8], Linfeng Zhang[1,‡]

[1]SJTU, [2]HKUST(GZ), [3]MBZUAI, [4]JLU, [5]THU, [6]HNU,
[7]Shanghai AI Lab, [8]UniTrento, [9]HKUST
[*]Equal Contribution [†]Project Lead [‡]Corresponding Author
cliao127@connect.hkust-gz.edu.cn
{haocong-he,zhanglinfeng}@sjtu.edu.cn

Dataset: https://huggingface.co/datasets/Eustia1/OmniCoT/
Model: https://huggingface.co/Eustia1/OmniCoT-R1
Code: https://github.com/Chenfei-Liao/OmniCoT
Page: https://chenfei-liao.github.io/OmniCoT.github.io/

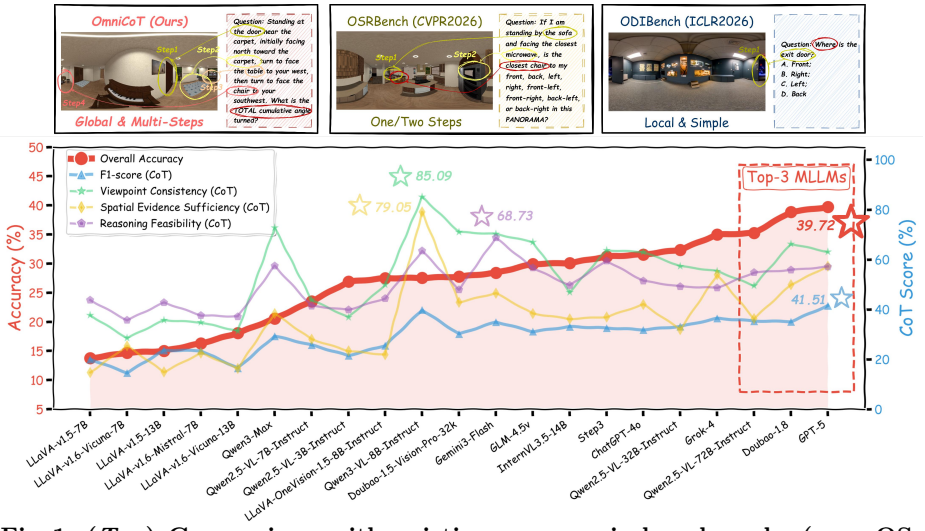


Fig. 1: (*Top*) Comparison with existing panoramic benchmarks (*e.g.*, OSRBench [8], ODIBench [37]). While prior works solely rely on synthetic data and primarily focus on few-step reasoning or local visual cues, OmniCoT demands comprehensive multi-hop inference across the full 360° view and uniquely incorporates a manually annotated real-world subset, driving MLLMs to fundamentally ***see more and reason more***. **(*Bottom*) Leaderboard of MLLMs on OmniCoT-B**. It showcases both the overall accuracy and detailed Chain-of-Thought (CoT) quality metrics.

Abstract. Multimodal Large Language Models (MLLMs) have demonstrated promising spatial reasoning capabilities, while these abilities remain underexplored in the emerging visual modality of panoramic imagery. The full 360° ×180° field of view of panoramas essentially supports complex global multi-step reasoning, which is also the fundamental advantage of panoramas in applications such as embodied intelligence. However, existing panoramic benchmarks largely focus on simplistic queries that rely on local cues or single-/few-step reasoning, thereby ignoring the fundamental advantage of panoramas and failing to fully exploit their potential. To address this gap, we introduce ***OmniCoT***, a panoramic spatial reasoning suite designed to enable MLLMs to use global evidence and perform multi-step inference across viewpoints. It includes ***OmniCoT-B*** (6.7K data) for evaluation, which measures both answer accuracy and reasoning quality, ***OmniCoT-Real*** (1K data) as a manually annotated real-world subset to quantify the Sim-to-Real gap. For training, ***OmniCoT-T*** (14.3K data) is purpose-built with structured stepwise Chain-of-Thought annotations that explicitly link intermediate reasoning steps to panoramic evidence. Based on OmniCoT-T, we introduce ***OmniCoT-R1*** and adopt a two-stage training strategy tailored to the geometrically complex panoramic space, where Supervised Fine-Tuning (SFT) anchors reasoning to panoramic evidence (e.g., bearings, proximity) and GRPO penalizes geometrically incoherent paths to consolidate global 360° spatial consistency. Through OmniCoT, we aim to ***recalibrate the difficulty of panoramic spatial reasoning to better align with the intrinsic capabilities of panoramic imagery***, thereby fostering meaningful progress in this research area.

Keywords: Panorama · Spatial Reasoning · Benchmark

1 Introduction

Panoramas, as an emerging visual modality, provide a unique 360°×180° field of view (FoV) for robots, vehicles, and related applications [17, 28, 41, 45]. By capturing more of the surrounding environment, this wide field of view is especially useful for tasks such as autonomous driving and embodied intelligence [36,38,41]. However, severe geometric distortions and dense visual content introduce substantial domain gaps, creating new challenges when extending MLLMs to understand panoramas. To evaluate the capability of MLLM in panoramic environments, VQA stands out as the most common way to test whether MLLMs can truly understand and reason about the entire 360° environment.

With more panoramic benchmarks proposed, a closer look shows a significant flaw, as shown in Fig. 1 and Table 1: most existing questions in current benchmarks remain overly simplistic, failing to compel state-of-the-art MLLMs [27,31, 32] to leverage the full 360° view. The currently proposed panoramic spatial reasoning benchmarks focus on only local regions (*e.g.*, "Where is the exit door"), and the reasoning tasks tend to require no more than two-hop inference. We believe that ***panoramas are built for complex spatial tasks***, because they

Table 1: Comparisons of our OmniCoT with other omnidirectional spatial reasoning benchmarks.

Benchmark	Spatial Scope	Annotation	Syn.	Real.	CoT	# of Q Types	# of Images	# of QA
VQA 360° [6]	Global/Single-step	Auto	✘	✔	✘	5	1.4k	16.9k
PanoEnv-QA [18]	Global/Single-step	Auto	✔	✘	✘	5	595	14.8k
OmniVQA [40]	Local/Single-step	Auto	✔	✘	✘	3	1.2k	4.8k
OSR-Bench [8]	Local/Few-step	Auto	✔	✘	✘	3	4.1k	153k
ODI-Bench [37]	Local/Few-step	Auto&Manual	✔	✔	✘	10	2k	4.3k
OmniCoT (Ours)	Global/Multi-hop	Auto&Manual	✔	✔	✔	6	4.2k	21.6k

show how every object in a full 360° view relates to the others, regardless of their distance or direction. When questions are too simple, they ignore the wealth of information in panoramas. To bridge this gap, we introduce ***OmniCoT***[1], which includes a challenging panoramic reasoning benchmark ***OmniCoT-B*** (6.7K data) that pushes models to ***"see more and reason more"***, a manually annotated real-world subset ***OmniCoT-Real*** (1K data) to quantify the Sim-to-Real gap, and a training set ***OmniCoT-T*** (14.3K data).

First, building upon real-world embodied perception requirements, we introduce a structured question taxonomy that decomposes panoramic spatial reasoning into three progressive dimensions: ***See-Locate-Move***. This taxonomy enables the systematic generation of multi-hop questions spanning viewpoint transformation, inter-object relational reasoning, and embodied action simulation in panoramic scenes. Then, we design a hybrid pipeline that integrates automated generation with expert supervision to construct high-quality question–answer pairs requiring multi-hop reasoning and global information. Based on OmniCoT-B, we perform extensive tests on a wide range of MLLMs, evaluating both their general and spatial reasoning abilities. Building upon the benchmark, we further develop OmniCoT-R1 based on OmniCoT-T through a two-stage training strategy, which serves as the baseline of panoramic reasoning MLLMs. The strategy first establishes structured reasoning via SFT, and then enhances the model's ability to solve complex multi-step tasks with GRPO. Moreover, we collect and annotate 1K real-world data to construct OmniCoT-Real, enabling the evaluation of the Sim-to-Real gap on OmniCoT.

At a glance, our contributions can be summarized as follows: ❶ We introduce ***OmniCoT-B***, a new benchmark that challenges MLLMs to fully use the 360° space. It moves beyond simple, local questions, requiring MLLMs to see more and reason more. ❷ We provide a thorough evaluation of current MLLMs, measuring not just if their final answers are correct, but whether their step-by-step reasoning actually makes sense. ❸ We develop ***OmniCoT-R1*** based on our proposed ***OmniCoT-T*** through a two-stage training strategy, which serves as the new baseline of the complex panoramic spatial reasoning task. ❹ We present

[1] Here "Omni" refers to panoramic (omni-directional) input rather than omni-modal input [35].

OmniCoT-Real, a set of real-world panoramas used to further measure the Sim-to-Real gap of OmniCoT. ❺ We provide diagnostic analyses in the Supplementary, including paraphrase robustness, input-projection robustness, tolerance sensitivity, human verification, and data scaling of OmniCoT-T.

2 Related Work

MLLMs for Omnidirectional Vision. While pinhole images only show a narrow slice of a scene, panoramas capture everything in a full 360° view [17]. This complete FoV is essential for embodied AI and virtual reality [44]. This makes panoramic imagery an ideal testing ground for MLLMs aiming to master higher-level spatial reasoning [8]. Current research in this area focuses on two main goals: ① building better datasets and ② adapting models for panoramic vision. For instance, datasets like OSR-Bench [8] and ODI-Bench [37] use a mix of MLLMs and human effort to create panoramic question-answering tasks. Meanwhile, researchers are fine-tuning models with techniques like GRPO to help them navigate the distorted, wide-angle world of panoramas: examples include 360-R1 [40] for spatial reasoning and ERP-RoPE [46] for better image processing. However, there is a catch: most of these models are still tested on very basic questions that only look at one small spot or require just one or two simple steps. This fails to use the full potential of panoramas and doesn't reflect how they would be used in the real world. Thus, we introduce OmniCoT, which is designed to calibrate the difficulty of panoramic spatial reasoning benchmarks by requiring MLLMs to see more and reason more.

MLLM Benchmarks for Spatial Reasoning. Spatial reasoning research has progressed from object recognition to goal-oriented tasks such as navigation and relational reasoning [7, 20, 39, 43]. These benchmarks evaluate how MLLMs understand the physical world by mapping surroundings and planning actions [5, 11, 25, 34]. However, a key gap remains in evaluating reasoning within 360° environments. Existing panoramic benchmarks often treat spatial reasoning as a "black box" [6, 8, 18, 37, 40], focusing only on final-answer correctness. This is inadequate for panoramas, where dense visual information enables models to exploit local shortcuts [1, 9] instead of true global reasoning. We argue that panoramic intelligence should be evaluated not only by outcomes but also by reasoning processes [16, 42]. While prior benchmarks emphasize accuracy, OmniCoT-T focuses on reasoning chains [15, 30], requiring step-by-step decomposition of thought processes to provide a more transparent and rigorous evaluation of spatial reasoning in 360° environments.

3 OmniCoT Design & Generation

3.1 Question Dimension: "See-Locate-Move"

Inspired by real-world embodied requirements, we design our complex panoramic spatial reasoning questions from three progressive dimensions: ***"see", "locate"***

and "move". The first dimension, "see", focuses on viewpoint transformation. It tests the MLLMs' ability to decouple their view from the camera's initial pose and multi-hop viewpoint reasoning, requiring them to mentally synthesize views from arbitrary angles and understand the spherical geometry of the panorama. The second dimension, "locate", advances to spatial object relationships. Questions of this dimension evaluate MLLMs' ability to reason about the inter-object spatial relationship, verifying whether it can perceive the scene not just as a bag of pixels, but as a structured layout of interconnected entities. The final dimension, "move", introduces embodied action simulation. This dimension challenges MLLMs to execute virtual movements and predict the visual consequences of these actions, thus evaluating their potential for real-world scenarios.

➢ **[See] Multi-hop Viewpoint Transformation.** This category assesses the MLLM's fundamental ability to observe the world through multi-step panoramas. This involves not only understanding spherical projections but also the ability to perform precise geometric integration. Specifically, ❶ Multi-Step Orientation Tracking (MOT) requires the model to start from an initial pose and execute a series of relative rotations, finally identifying the target object. ❷ Relative Angular Calculation (RAC) challenges the model to compute the cumulative angular displacement or the specific bearing between multiple landmarks, treating objects as dynamic reference points. These tasks verify if the model can maintain spatial constancy even when the target undergoes significant perspective warping near the panoramic poles or across the image boundaries.

Type A1 Example: *Standing at the [object_desc], facing [cardinal direction], turn [angle1]°[dir1], then turn [angle2]°[dir2],* ***what is the NEAREST object?***
Type A2 Example: *Standing at [object_desc], initially facing [object_desc], turn to face [object_desc], then turn to face [object_desc].* ***What is the TOTAL cumulative angle turned?***

➢ **[Locate] Inter-Object Spatial Relationship.** This category assesses the MLLM's ability to reason about the spatial relationships between multiple objects within a scene through multi-hop reasoning. Rather than answering questions based on direct observation, the model must navigate chains of spatial relations and use qualifiers such as "nearest" or "second closest" to either identify a specific target object or determine its directional relationship to an anchor object. Specifically, ❶ Multi-Hop Object Identification (MOI) requires the model to sequentially apply spatial qualifiers across multiple reference objects to find a unique target. ❷ Multi-Hop Direction Identification (MDI) challenges the model to determine the final direction (*e.g.*, North) of a target relative to an anchor.

Type B1 Example: *What [object_desc] is directly to the [cardinal direction] of the NEAREST object that is [cardinal direction] of the [object_desc]?*
Type B2 Example: *In which direction is the [object_desc] to the [cardinal direction] of the [object_desc], relative to the [object_desc] itself?*

➢ **[Move] Embodied Action Simulation.** Furthermore, to evaluate the MLLM's ability to perform dynamic spatial reasoning through the simulation of embodied actions, such as movement and turning, the questions of Type C are introduced. Within this category, the model must track continuous state changes including updates to position, orientation, and field of view, for either

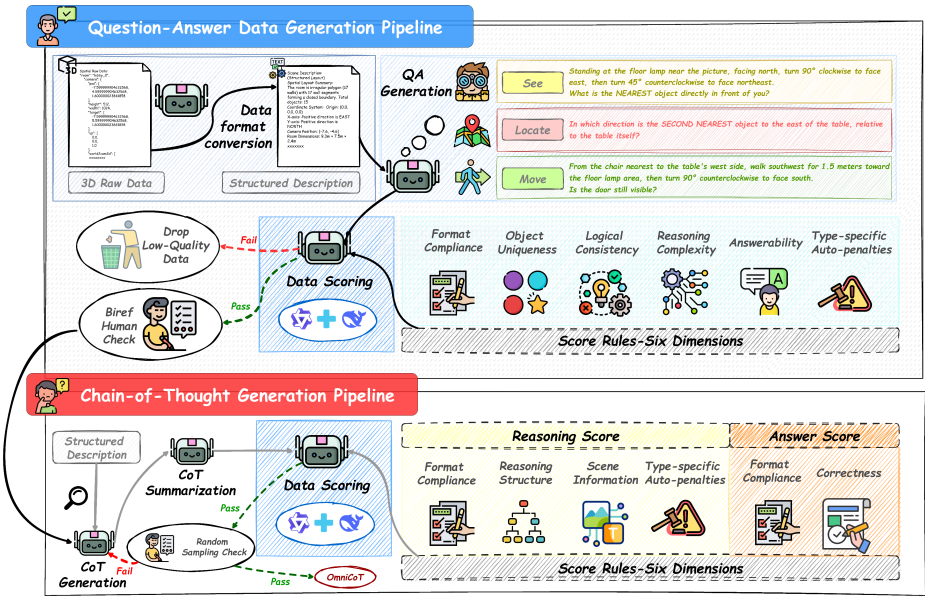


Fig. 2: Data Generation Pipeline of OmniCoT. It translates 3D scene geometry into structured language representation, generates multidimensional candidate questions (See, Locate, Move), and applies rigorous dual-LLM reasoning judges to select high-quality QA pairs. Finally, a structured three-stage process synthesizes, refines, and evaluates the corresponding step-by-step Chain-of-Thought (CoT) rationales, culminating in a manually verified dataset.

itself or objects within the scene. Specifically, ❶ Pure Translational Movement (PTM) simulates navigation along a single direction without rotation, testing the model's capacity to predict objects encountered along a linear path. ❷ Rotation-Translational Movement (RTM) combines straight-line movement with a subsequent rotation, requiring the model to update its egocentric perspective and reason about the altered spatial relationships and visibility.

Type C1 Example: *From the [object_desc], near the [object_desc],, walk straight [cardinal direction] for [number] meters.* ***What is the FIRST object you will encounter?***
Type C2 Example: *From the [object_desc],, walk [cardinal direction] [number] meters toward the [object_desc], area, then turn [angle] to face [cardinal direction].* ***Is the [object_desc], still visible from your new position and facing direction?***

3.2 Data Generation

Step 1: Question-Answer (QA) Generation. In the process of QA generation shown in Fig. 2, we propose a hierarchical generation pipeline to gradually generate and optimize the panoramic spatial reasoning QA pairs. The generation process begins with reliable raw 3D scene data, which is transformed into structured natural language descriptions. This representation provides rich and

reliable spatial context for subsequent question generation and enables questions that fully leverage the global perspective afforded by panoramas. Building on this representation, we generate multiple batches of candidate questions covering all predefined question types. We then employ two reasoning-focused large language models (DeepSeekv3.2 [19] and Qwen3-Max) as judges to score these candidates from six dimensions: format compliance, object uniqueness, logical consistency, reasoning complexity, answerability, and special type-specific auto-penalties. Only candidate questions that meet the scoring criteria will proceed to the final manual review. Experts will quickly check the generated question-answer pairs to ensure there are no significant deviations in format or content.

Step 2: Chain-of-Thought Generation. Building upon the validated QA pairs, we implement a structured three-stage CoT generation pipeline to produce step-by-step spatial reasoning traces as shown in Fig. 2. Firstly, we utilize a specialized prompting framework that adapts to each question type's unique reasoning requirements, guaranteeing that each generated CoT must demonstrate 2-4 clear reasoning steps, use appropriate transition words, and maintain strict natural language compliance throughout. Then, the summarizing process distills the core logical flow, preserving critical spatial information while eliminating redundancies. Finally, we implement a rigorous scoring mechanism evaluating both the reasoning process and answer quality across six critical dimensions: reasoning format compliance, reasoning structure, scene information utilization, type-specific auto-penalties, answer format compliance, and answer correctness. The QA pairs with the passed CoT will go through the final expert check process. Considering the over-substantial time and cognitive effort required for the careful assessment of CoT data, we randomly sampled near 400 data samples for experts to conduct rigorous and detailed evaluations. Until the random evaluation result satisfies the accuracy requirements of 95%, the final results will be accepted as the final version of OmniCoT. The specific random evaluation details are presented in the ***Supplementary*** and statistics are shown in Fig. 3

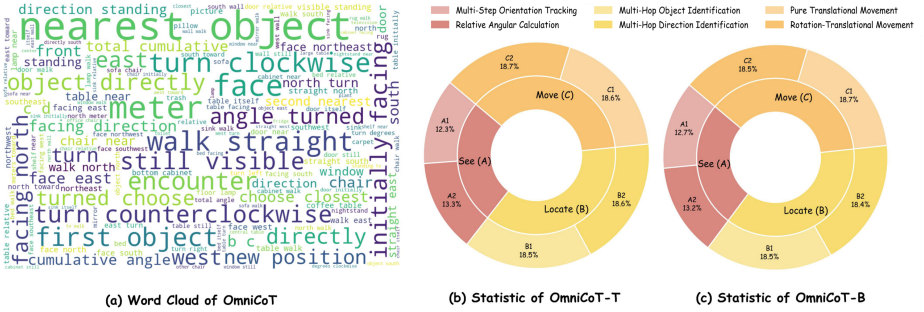
4 Benchmark

4.1 Evaluation Design

To assess spatial reasoning CoT on OmniCoT-B, we adopt a two-dimensional framework with six metrics capturing both general and spatial reasoning quality.

General Reasoning Quality. Following MME-CoT [15], we employ three interpretable metrics to assess the general reasoning quality: Precision, Recall, and F1-score. Precision measures the faithfulness of each generated step by calculating the ratio of correct steps. Recall evaluates reasoning informativeness by quantifying the proportion of ground-truth solution steps that appear in the model's response. Moreover, F1-score provides a balanced overall assessment. The evaluation process and prompts of judge LLMs for this part are strictly consistent with MME-CoT [15].

Spatial Reasoning Quality. We introduce three fine-grained metrics specifically for evaluating spatial reasoning quality in panoramas: Viewpoint Consis-



(a) Word Cloud of OmniCoT (b) Statistic of OmniCoT-T (c) Statistic of OmniCoT-B

Fig. 3: OmniCoT exhibits broad lexical coverage and balanced question-type composition. (a) Word cloud over all OmniCoT questions. **(b)** Question-type distribution of OmniCoT-T shows a balanced mixture of *See–Locate–Move* taxonomy. **(c)** Question-type distribution of OmniCoT-B, which closely matches OmniCoT-T and thus mitigates train–test distribution shift.

tency (VC), Spatial Evidence Sufficiency (SES), and Reasoning Feasibility (RF). VC assesses model's ability to maintain coherent spatial perspectives throughout multi-hop reasoning. All viewpoint-related statements are extracted and scored individually against ground-truth geometry, with the final score being the average consistency. SES measures whether the reasoning process adequately cites necessary spatial relationships. Key relationships required to answer the question are identified and checked for citation, with the citation rate serving as the final score. RF evaluates the physical plausibility of execution steps (movement, turning, interaction) within the scene constraints. Each execution step is scored based on feasibility, with the average determining the final score.

4.2 Experimental Settings & Results

In the evaluation process, we benchmark 12 open-source MLLMs (including Qwen2.5-VL [4], Qwen3-VL [3], LLaVA-OneVision [2], LLaVA-v1.5 [21], LLaVA-v1.6 [22], and InternVL3.5 [29]) and 9 closed-source MLLMs (including ChatGPT-4o [14], GPT-5 [24], Doubao-1.5 [10], Doubao-1.8 [23], Gemini3-Flash [26], GLM-4.5v [12], Step3 [13], Qwen3-Max [3], and Grok-4 [33]) on the OmniCoT-B. The input consists of an ERP panoramic image along with a corresponding question, supplemented with the coordinates and orientation indications (e.g., positive X-axis pointing east, positive Y-axis pointing north) of three randomly selected objects in the scene to establish a unified spatial reference frame. We refer to these coordinate-and-orientation hints as *spatial anchors*. A spatial anchor is an object-level reference point with known position and orientation, used to align the visual panorama with the textual spatial description. *Anchor density* denotes the number of such reference objects provided in the prompt. We use 3 anchors by default because they provide a lightweight global frame while avoiding exposure of the full scene layout, and random selection prevents the evaluation from favoring object-specific priors. To disentangle panoramic reasoning from ERP

Table 2: Performance of SoTA MLLMs on OmniCoT-B. Accuracy is evaluated on three spatial reasoning dimensions: ***See*** (Multi-hop Viewpoint Transformation), ***Locate*** (Inter-object Spatial Relations), and ***Move*** (Embodied Action Simulation). CoT quality is measured by general metrics [15]: ***Pre.*** (Precision), ***Rec.*** (Recall), and ***F1***, along with panorama-specific metrics: ***VC*** (Viewpoint Consistency), ***SES*** (Spatial Evidence Sufficiency), and ***RF*** (Reasoning Feasibility). Best and worst results are highlighted in green and red, respectively.

Model	*Accuracy Evaluation*				*Chain-of-Thought Evaluation*					
	See	***Locate***	***Move***	***Overall***	***Pre.***	***Rec.***	***F1.***	***VC***	***SES***	***RF***
Open-Source MLLMs										
Qwen2.5-VL-3B-Instruct [4]	34.27	17.15	31.43	26.90	26.17	18.23	21.49	36.85	23.36	39.82
Qwen2.5-VL-7B-Instruct [4]	33.77	15.53	24.33	23.53	34.60	20.52	25.76	43.85	28.05	41.40
Qwen2.5-VL-32B-Instruct [4]	47.84	21.93	31.99	32.39	42.05	27.39	33.17	57.33	31.97	49.18
Qwen2.5-VL-72B-Instruct [4]	52.03	23.69	35.05	35.26	46.57	28.53	35.39	49.51	36.30	54.80
Qwen3-VL-8B-Instruct [3]	41.52	18.73	26.59	27.56	61.22	29.48	39.80	85.09	79.05	63.49
LLaVA-OneVision-1.5-8B-Instruct [2]	42.40	14.61	29.95	27.51	36.99	19.53	25.56	49.88	21.84	44.43
LLaVA-v1.5-7B [21]	15.00	9.82	16.80	13.76	30.56	12.93	20.04	37.66	14.72	43.76
LLaVA-v1.5-13B [21]	21.01	14.17	11.71	15.02	40.88	16.60	23.62	35.70	14.93	42.76
LLaVA-v1.6-Mistral-7B [22]	22.01	17.24	11.40	16.31	35.18	17.64	23.49	34.80	22.56	37.64
LLaVA-v1.6-Vicuna-7B [22]	11.13	12.63	19.24	14.70	30.56	9.60	14.61	28.52	25.41	35.66
LLaVA-v1.6-Vicuna-13B [22]	16.32	15.00	22.33	18.07	30.40	11.31	16.48	31.36	16.41	37.12
InternVL3.5-14B [29]	41.77	20.35	31.64	30.11	40.54	27.99	33.11	46.94	36.18	49.61
Close-Source MLLMs										
ChatGPT-4o [14]	51.97	15.18	33.65	31.58	42.46	25.51	31.87	62.99	42.03	51.41
GPT-5 [24]	57.29	24.40	42.71	39.72	51.43	34.79	41.51	63.05	57.30	57.10
Doubao-1.5-Vision-Pro-32k [10]	54.91	8.38	28.08	27.77	45.47	22.53	30.13	71.02	42.85	47.95
Doubao-1.8 [23]	57.09	23.87	41.14	38.90	42.56	29.92	35.13	66.29	49.94	55.79
Gemini3-Flash [26]	47.52	14.17	29.29	28.43	79.30	22.46	35.01	70.22	46.54	68.73
GLM-4.5v [12]	53.15	15.05	28.60	29.95	47.82	23.07	31.12	66.98	38.31	56.63
Step3 [13]	48.40	16.71	33.69	31.23	58.59	22.53	32.55	63.71	36.96	59.49
Grok-4 [33]	54.40	19.30	37.04	34.99	48.28	29.21	36.40	55.36	53.88	48.65
Qwen3-Max [3]	52.03	6.19	12.97	20.58	43.88	22.03	29.33	72.76	38.39	57.48

distortion, we further evaluate multi-view perspective inputs in the Supplementary. DeepSeekv3.2 [19] is selected as the judge model for CoT scoring.

As summarized in Table 2, our evaluation reveals divergent performance patterns across models, where closed-source titans like GPT-5 and Doubao-1.8 establish a stronghold in the ***Move*** dimension (over 41% accuracy), demonstrating superior stability in ego-perspective updates. While large-scale models such as Qwen2.5-VL-72B excel in geometric projection (***See***), a universal bottleneck persists in the ***Locate*** dimension, where no model surpasses the 25% threshold, highlighting the difficulty of multi-hop topological reasoning. Interestingly, fine-grained CoT metrics expose a precision-informativeness trade-off: high-scale models like Gemini3-Flash achieve near-perfect Precision (79.30%) but suffer from sparse reasoning traces (low Recall), whereas smaller models like Qwen3-VL-8B exhibit higher fidelity in spatial anchoring (VC and SES). These results underscore that while scaling enhances raw perception, it does not solve the challenge of consistent environmental grounding in complex panoramic scenes.

5 Extensive Results

5.1 CoT *v.s.* Direct Answer

Divergent Impact of CoT on MLLMs: To further investigate whether Chain-of-Thought (CoT) genuinely enhances spatial reasoning in panoramas, we conduct a controlled comparative experiment. For each model, we evaluate under two distinct response modes: ❶ CoT, ❷ Direct Answer, with the results shown in Table 3. For open-source models, CoT acts as an effective panoramic spatial navigator. By externalizing multi-hop logic, open-source models' panoramic reasoning ability is improved. Conversely, forcing CoT on closed-source models (*e.g.*, Gemini3-Flash) actively degrades their accuracy. Forcing them to write out a step-by-step reasoning path is proven to be unnecessary and tends to hurt their performance. This divergent impact suggests that ***the verbose generation of long reasoning steps inherently carries a risk of cumulative hallucination, which currently outweighs the logical benefits of CoT for top-tier closed-source models.***

Table 3: Impact of CoT on Overall Accuracy. We compare **w/o CoT** (direct answering) with **w/ CoT** (thinking). Δ denotes absolute accuracy shift (**w/CoT** *v.s.* **w/o CoT**). RED and BLUE indicate gains and drops.

Model	w/o CoT	w/ CoT	Δ
Open-Source MLLMs			
Qwen3-VL-8B-Instruct [3]	25.39	**27.56**	+2.17
Qwen2.5-VL-3B-Instruct [4]	21.20	**26.90**	+5.70
Qwen2.5-VL-7B-Instruct [4]	23.01	**23.53**	+0.52
Qwen2.5-VL-72B-Instruct [4]	26.51	**35.26**	+8.75
InternVL3.5-14B [29]	23.79	**30.11**	+6.32
LLaVA-v1.5-7B [21]	4.91	**13.77**	+8.86
LLaVA-v1.5-13B [21]	17.38	**15.03**	-2.35
LLaVA-v1.6-Mistral-7B [22]	16.70	**16.31**	-0.39
LLaVA-v1.6-Vicuna-7B [22]	9.77	**14.70**	+4.93
LLaVA-v1.6-Vicuna-13B [22]	18.52	**18.07**	-0.45
Close-Source / API MLLMs			
Qwen3-Max [3]	25.39	**20.58**	-4.81
GPT-5 [24]	42.66	**39.72**	-2.94
Doubao-1.5-Vision-Pro-32k [10]	30.96	**27.76**	-3.20
Gemini3-Flash [26]	39.97	**28.44**	-11.53

Coupling between Reasoning Quality and Accuracy: We further compare the trends of four CoT quality metrics (F1., VC, SES, RF) and the overall accuracy, which are visualized in Fig. 1. Overall, accuracy exhibits a positive directional synergy with spatial reasoning metrics, confirming the intuition that better reasoning generally yields better predictions. However, we also observe a paradoxical "Reasoning-Answer Mismatch" in certain outliers. A striking example is Qwen3-VL-8B-Instruct: it achieves relatively high RF, VC, and SES, but fails to deliver a commensurate final accuracy. This exposes a phenomenon we term "linguistic blind-navigation". The model is highly articulate, capable of writing out a physically flawless, step-by-step navigation plan (*e.g.*, "turn 90 degrees right, walk past the sofa"). Yet, this textual logic is disconnected from the actual visual geometry. Because panoramic images suffer from severe spherical distortions, the model's pure language skills create a "persuasive illusion" of reasoning, but it ultimately fails to map those perfectly written textual steps back to the distorted pixels to locate the correct target. Ultimately, the "linguistic blind-navigation" phenomenon reveals a critical flaw in some MLLMs: ***their internal reasoning process is often driven by text priors rather than true cross-modal spatial perception.***

5.2 What Makes Better Reasoning in Panoramas

Scaling Laws in Panoramic Spatial Reasoning: We investigate the scaling behavior of panoramic spatial reasoning using the Qwen2.5-VL family (3B, 7B, 32B, and 72B). As shown in Fig. 4, we observe a clear scaling dividend in overall accuracy, which improves steadily from 26.90% (3B) to 35.26% (72B). However, this scaling trajectory exhibits a notable marginal effect. The performance leap from 7B to 32B is substantial, yet the subsequent push to 72B yields only minor incremental gains. Similarly, the scaling of reasoning quality metrics follows a non-linear pattern, peaking early and plateauing at larger scales. This indicates that ***while increasing the parameter count successfully broadens the MLLM's ability for handling panoramas, simple scaling eventually hits a bottleneck.*** To further break through the performance ceiling in panoramas, relying solely on larger models is insufficient; explicitly optimizing the model's spatial reasoning processes (*e.g.*, Post-training) becomes a more promising path.

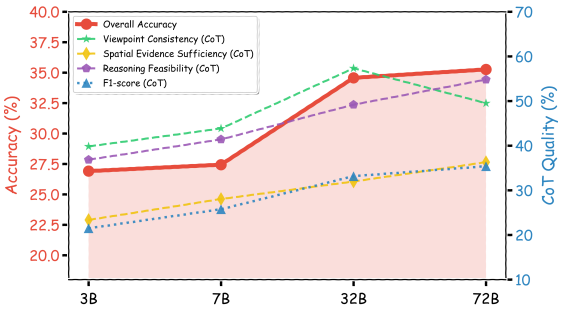


Fig. 4: Scaling Analysis of Panoramic Spatial Reasoning. We evaluate the Qwen2.5-VL model family across varying parameter scales (3B to 72B).

The Impact of Spatial Anchor Density: To understand the role of the environmental grounding anchor, we evaluate model performance under varying densities of reference points (0, 3, and 10). As summarized in Table 4 and Fig. 5, increasing the number of spatial anchors yields a "precision dividend" across most MLLMs. These reference points serve as global coordinate "pins" that mitigate spatial drift. For instance, InternVL3.5-14B achieves a substantial 5.88% accuracy boost when anchors are increased from 0 to 10. Interestingly, we observe that providing even a few anchors significantly activates the latent spatial reasoning capabilities of top-tier models like GPT-5 at the 10-point setting. This confirms that ***explicit spatial anchors act as crucial global co-***

Table 4: Performance comparison of selected models with different numbers of initial reference points (0, 3, 10).

Points	*Accuracy Evaluation*				*Chain-of-Thought Evaluation*					
	See	*Locate*	*Move*	*Overall*	*Pre.*	*Rec.*	*F1.*	*VC*	*SES*	*RF*
Qwen2.5-VL-7B-Instruct [4]										
0	30.70	22.37	25.77	25.79	40.25	22.2	28.96	38.95	35.13	42.18
3	33.77	15.53	24.33	23.53	34.60	20.52	25.76	43.85	28.05	41.40
10	35.33	22.07	33.17	29.63	48.89	29.79	37.02	48.79	37.68	48.79
LLaVA-v1.6-Vicuna-7B [22]										
0	18.26	14.87	20.98	18.02	40.49	15.25	22.16	33.18	15.03	38.65
3	11.13	12.63	19.24	14.70	30.56	9.60	14.61	28.52	25.41	35.66
10	12.75	13.99	18.80	15.46	32.76	14.88	20.46	30.06	16.06	37.97
InternVL3.5-14B [29]										
0	36.64	19.26	31.60	28.35	41.71	27.67	33.27	40.72	35.14	45.66
3	41.77	20.35	31.64	30.10	40.54	27.99	33.11	46.94	36.18	49.61
10	44.34	25.23	36.13	34.23	44.43	29.58	35.52	54.59	43.40	55.52
GPT-5 [24]										
0	56.47	28.21	41.84	40.60	50.58	35.48	41.71	54.88	50.67	54.75
3	57.29	24.40	42.71	39.72	51.43	34.79	41.51	63.05	57.30	57.10
10	58.66	29.74	48.01	44.03	64.80	42.62	51.42	77.98	76.96	71.18
Gemini3-Flash [26]										
0	47.28	24.97	32.13	33.41	69.18	24.13	35.78	53.26	36.77	59.01
3	47.52	14.17	29.29	28.43	79.30	22.46	35.01	70.22	46.54	68.73
10	53.28	31.11	46.36	42.52	81.49	30.64	44.53	77.19	71.42	79.87

ordinates that effectively mitigate spatial drift and activate the latent reasoning capabilities of MLLMs.

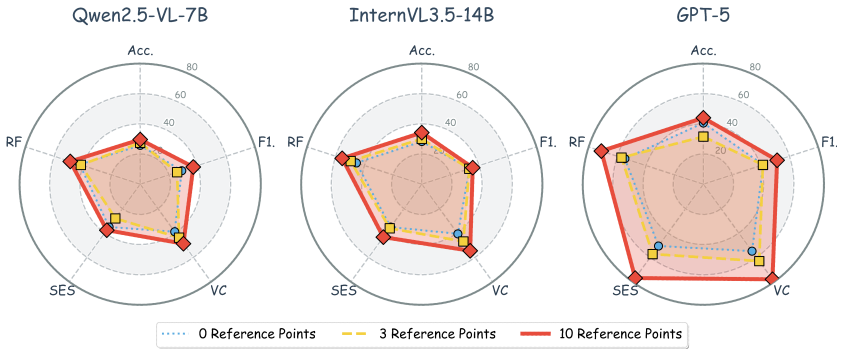


Fig. 5: Capability Expansion via Spatial Anchors. Radar charts illustrate model performance across Accuracy and four CoT metrics with varying numbers of reference points (0, 3, 10).

5.3 Small-Scale Real-world Test

To evaluate MLLMs' panoramic spatial reasoning capability in real-world environments and quantify the gap between synthetic and real-world settings, we construct ***OmniCoT-Real***. Specifically, we collect 200 real-world panoramas with Insta-X5, spanning 13 diverse indoor scene categories, including living rooms, kitchens, bathrooms, corridors, classrooms, and gyms, among others. All QA pairs are manually annotated by three domain experts over approximately 320 hours of collaborative effort, yielding 1,073 QA pairs in total. Compared to OmniCoT-B, grounded in simulated 3D scenes with precise coordinates, OmniCoT-Real introduces authentic real-world challenges, including lens distortion, uneven lighting, and object occlusion, making it a more demanding testbed for evaluating generalization.

Table 5: Evaluation on OmniCoT-Real, showing performance comparison on real-world panoramas.

Model	*Accuracy*				*CoT*		
	See	*Locate*	*Move*	*Overall*	*Pre.*	*Rec.*	*F1.*
Qwen2.5-VL-7B-Instruct [4]	18.48	5.74	20.59	15.28	35.20	12.53	18.48
LLaVA-v1.6-Vicuna-7B [22]	11.74	15.51	04.23	11.74	25.00	7.19	11.17
InternVL3.5-14B [29]	25.54	8.70	32.35	22.73	40.52	17.01	23.97
GPT-5 [24]	30.71	17.22	39.04	29.45	55.69	25.31	34.80
Gemini3-Flash [26]	33.15	15.11	40.37	30.10	58.61	32.19	41.56

As illustrated in Table 5 and Fig. 6, transitioning from synthetic environments to real-world panoramas exposes a substantial sim-to-real performance gap. The majority of evaluated MLLMs suffer significant declines in both overall accuracy and spatial reasoning quality (F1-score) on OmniCoT-Real. Specifically, all models consistently score lowest on the *Locate* dimension, suggesting that fine-grained inter-object relationship reasoning is particularly

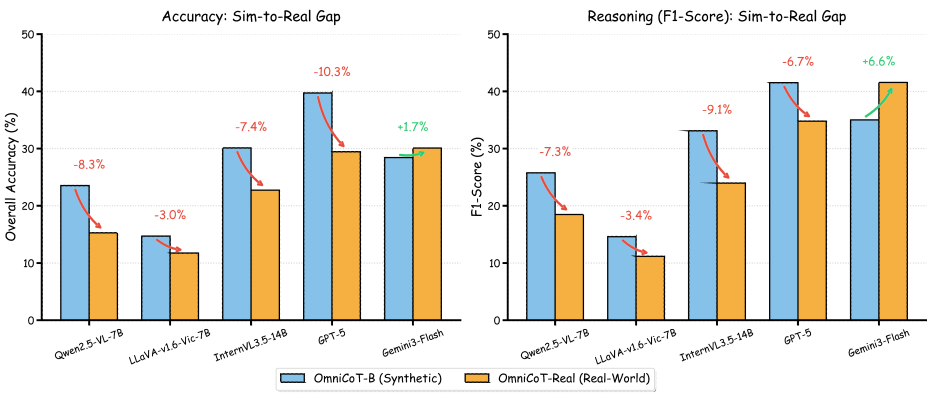


Fig. 6: Sim-to-Real Performance Gap. Performance comparison between the synthetic benchmark (OmniCoT-B) and the real-world evaluation set (OmniCoT-Real). We evaluate both Overall Accuracy (left) and Spatial Reasoning Quality via F1-Score (right). A clear performance degradation is observed when transitioning from simulated scenes to real-world continuous environments, highlighting the persistent challenge of real-world generalization for MLLMs in panoramic environments.

vulnerable to real-world visual complexity. Furthermore, open-source models exhibit notably low Recall compared to their closed-source counterparts, reflecting a failure to ground sufficient spatial evidence when confronted with real-world panoramas. Ultimately, the stark contrast between synthetic and real-world performance underscores that simulated benchmarks alone are insufficient. This serves as a compelling call to action, emphasizing ***the indispensable role of large-scale, real-world panoramic data in developing and evaluating truly robust, generalizable MLLMs.***

5.4 OmniCoT-R1

Effectiveness of SFT and GRPO: Building on our proposed OmniCoT-T, we develop ***OmniCoT-R1*** by post-training Qwen2.5-VL with a two-stage strategy including SFT and GRPO. During the SFT stage, we explicitly instruct the model to follow a structured panoramic reasoning protocol, thereby yielding an explicit `<think>...<answer>` format where each reasoning step is strictly grounded in visual evidence from the panorama. Initialized from this SFT checkpoint, we subsequently employ GRPO to further refine the model's long-

Table 6: Implementation details of OmniCoT-R1, including key hyperparameters for SFT and GRPO.

Item	Setting
Base model	Qwen2.5-VL-7B-Instruct [4]
Trainable modules (SFT)	Language model only, vision+proj frozen
Trainable modules (GRPO)	Full model fine-tuning
Precision	BF16
Max sequence length	4096
Max completion length	2048
SFT data	OmniCoT-T
SFT LR schedule	Cosine, warmup 0.03
GRPO data	OmniCoT-T
GRPO generations	8
Rewards	format / accuracy / repetition
Reward weights	0.1 / 1.0 / 0.2
GRPO learning rate	1e-6

Table 7: Results of OmniCoT-R1. Performance evaluation on OmniCoT-B comparing the baseline Qwen2.5-VL-7B-Instruct against our models trained via Supervised Fine-Tuning (SFT) and subsequent Group Relative Policy Optimization (GRPO). Both accuracy and spatial reasoning metrics show consistent gains through the two-stage post-training pipeline.

Model	*Accuracy Evaluation*				*Chain-of-Thought Evaluation*					
	See	*Locate*	*Move*	*Overall*	*Pre.*	*Rec.*	*F1.*	*VC*	*SES*	*RF*
Qwen2.5-VL-7B-Instruct [4]	33.77	15.53	24.33	23.53	34.60	20.52	25.76	43.85	28.05	41.40
OmniCoT-R1 (SFT)	59.41	39.44	52.06	49.31	42.89	34.22	38.07	46.25	57.87	53.93
OmniCoT-R1 (SFT+GRPO)	67.97	51.82	61.34	59.54	55.81	47.19	51.14	59.20	66.64	63.21

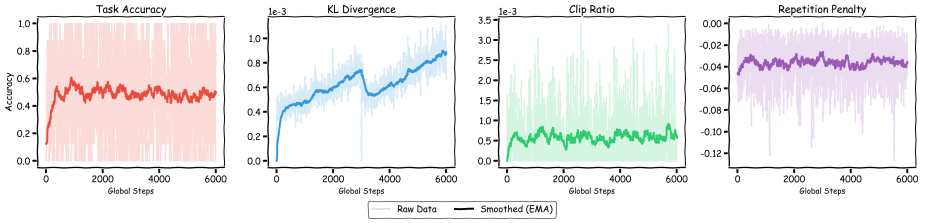


Fig. 7: Training Dynamics of OmniCoT-R1 (GRPO Stage). We track the evolution of four key metrics over the initial 6,000 optimization steps. Raw step-wise data is shown as a faded background, overlaid with an Exponential Moving Average (EMA) trendline for clarity.

horizon spatial reasoning capabilities. The reward combines format compliance to enforce the structured output schema, task accuracy computed by our geometry-grounded executor for OmniCoT, and anti-degeneration regularization to discourage verbose or repetitive chains. The details of the implementation are summarized in Table 6 and training dynamics of OmniCoT-R1 is shown in Fig. 7.

Training Stability: Beyond accuracy, training dynamics exhibit consistently healthy behavior. The ThinkAnswerFormat reward reaches a perfect score from the very first GRPO step and remains saturated throughout training, confirming that SFT reliably establishes the structural prerequisite for reward-driven optimization. The Kullback-Leibler (KL) divergence between the updated policy and the reference model increases only marginally and stabilizes within a narrow bound, indicating that the model acquires enhanced spatial reasoning capabilities without undergoing catastrophic forgetting or deviating significantly from the base policy. Furthermore, the surrogate objective clipping ratios consistently remain near zero, ensuring that policy updates are strictly confined within the GRPO trust region. The Repetition Penalty signal exhibits steady improvement over the training horizon, demonstrating a reduction in output degeneration and promotes the generation of more diverse, non-repetitive reasoning traces.

Collectively, these results provide concrete evidence for the effectiveness of OmniCoT-T as a post-training resource: it drives consistent reward improvements, maintains stable optimization dynamics, and reduces degeneration over long-horizon updates. Overall, OmniCoT-R1 serves as a successful baseline and pipeline for the panoramic spatial reasoning task, providing the community with a validated data foundation and a reproducible Reinforce-Learning paradigm for future panoramic spatial reasoning research.

6 Conclusion

In this paper, we identify a fundamental limitation in existing panoramic spatial reasoning benchmarks: the tendency to focus on simplistic, few-hops questions that fail to exploit the native advantage of panoramas. To address this, we introduce ***OmniCoT***, a suite encompassing the ***OmniCoT-B*** benchmark, the ***OmniCoT-Real*** subset, which enables the assessment of sim-to-real generalization, and the ***OmniCoT-T*** training dataset, explicitly designed to demand global and multi-hop reasoning across three progressive dimensions: ***See***, ***Locate***, and ***Move***. Our experiments reveal that CoT yields divergent effects across model families, and that panoramic reasoning quality is jointly shaped by model scale and explicit spatial grounding, with anchor density proving a particularly effective lever for activating latent reasoning capabilities. Moving beyond synthetic scenes, the substantial sim-to-real gap observed on OmniCoT-Real further highlights the necessity of real-world panoramic data and more faithful evaluation protocols for robust generalization. In conclusion, OmniCoT recalibrates the difficulty of panoramic reasoning benchmarks and provides a unified testbed for diagnosing both final-answer accuracy and intermediate reasoning quality. We hope this benchmark will facilitate principled comparisons across MLLMs, inspire stronger panorama-specific grounding and planning strategies. We will release the dataset, evaluation toolkit, and baseline models to support reproducible research and encourage future extensions to broader interaction settings.

Acknowledgements

This work was supported by the EU Horizon projects ELIAS (No. 101120237) and ELLIOT (No. 101214398) and by the FIS project GUIDANCE (No. FIS2023-03251).

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4971–4980 (2018)
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
6. Chou, S.H., Chao, W.L., Lai, W.S., Sun, M., Yang, M.H.: Visual question answering on 360deg images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1607–1616 (2020)
7. Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., et al.: Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7395–7408 (2025)
8. Dongfang, Z., Zheng, X., Weng, Z., Lyu, Y., Paudel, D.P., Van Gool, L., Yang, K., Hu, X.: Are multimodal large language models ready for omnidirectional spatial reasoning? arXiv preprint arXiv:2505.11907 (2025)
9. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
10. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
11. Guo, X., Zhang, R., Duan, Y., He, Y., Nie, D., Huang, W., Zhang, C., Liu, S., Zhao, H., Chen, L.: SURDS: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. arXiv preprint arXiv:2411.13112 (2024)
12. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
13. Huang, A., Yao, C., Han, C., Wan, F., Guo, H., Lv, H., Zhou, H., Wang, J., Zhou, J., Sun, J., et al.: Step3-vl-10b technical report. arXiv preprint arXiv:2601.09668 (2026)
14. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
15. Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., et al.: Mme-cot: Benchmarking chain-of-thought in large mul-

timodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621 (2025)
16. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: The twelfth international conference on learning representations (2023)
17. Lin, X., Ge, X., Zhang, D., Wan, Z., Wang, X., Li, X., Jiang, W., Du, B., Tao, D., Yang, M.H., et al.: One flight over the gap: A survey from perspective to panoramic vision. arXiv preprint arXiv:2509.04444 (2025)
18. Lin, Z., Zheng, X.: Panoenv: Exploring 3d spatial intelligence in panoramic environments with reinforcement learning. arXiv preprint arXiv:2602.21992 (2026)
19. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: Deepseek-v3. 2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556 (2025)
20. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
22. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge (2024)
23. Seed, B.: Seed1. 8 model card: Towards generalized real-world agency. Tech. rep., Technical report (model card), December 2025. URL https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/research/Seed-1.8-Modelcard.pdf (2025)
24. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
25. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Benchmarking spatial reasoning in vision-language models. arXiv preprint arXiv:2503.19707 (2025)
26. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
27. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al.: Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
28. Wang, C., Lin, X., Liu, J., Liu, Y., Wang, Z., Qi, D., Yan, Y., Chen, X.: Panoworld: Towards spatial supersensing in 360° panorama world. arXiv preprint arXiv:2605.13169 (2026)
29. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
30. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
31. Wen, Z., Wang, S., Zhou, Y., Zhang, J., Zhang, Q., Gao, Y., Chen, Z., Wang, B., Li, W., He, C., et al.: Efficient multi-modal large language models via progressive consistency distillation. arXiv preprint arXiv:2510.00515 (2025)
32. Wen, Z., Yang, B., Chen, S., Zhang, Y., Han, Y., Ke, J., Wang, C., Fu, Y., Zhao, J., Yao, J., et al.: Innovator-vl: A multimodal large language model for scientific discovery. arXiv preprint arXiv:2601.19325 (2026)

33. xAI: Grok 4 Model Card. Tech. rep. (Aug 2025), https://data.x.ai/2025-08-20-grok-4-model-card.pdf, model card / technical report
34. Xu, P., Wang, S., Zhu, Y., Li, J., Zhang, Y.: SpatialBench: Benchmarking multimodal large language models for spatial cognition. arXiv preprint arXiv:2511.21471 (2025)
35. Yagi, Y.: Omnidirectional sensing and its applications. IEICE transactions on information and systems **82**(3), 568–579 (1999)
36. Yang, J., Chen, Y., Xu, Y., Li, P., Wu, X., Wen, Z., Fang, B., Yu, T., Zhang, Z., Li, Y., et al.: Uaor: Uncertainty-aware observation reinjection for vision-language-action models. arXiv preprint arXiv:2602.18020 (2026)
37. Yang, L., Duan, H., Tao, R., Cheng, J., Wu, S., Li, Y., Liu, J., Min, X., Zhai, G.: ODI-Bench: Can MLLMs understand immersive omnidirectional environments? arXiv preprint arXiv:2510.11549 (2025)
38. Yang, Y., Wang, Y., Wen, Z., Zhongwei, L., Zou, C., Zhang, Z., Wen, C., Zhang, L.: Efficientvla: Training-free acceleration and compression for vision-language-action models. arXiv preprint arXiv:2506.10100 (2025)
39. Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., Wu, Y., Cui, R., Ran, B., Zhang, Z., et al.: How far are vlms from visual spatial intelligence? a benchmark-driven perspective. arXiv preprint arXiv:2509.18905 (2025)
40. Zhang, X., Ye, Z., Zheng, X.: Towards omnidirectional reasoning with 360-r1: A dataset, benchmark, and grpo-based method. arXiv preprint arXiv:2505.14197 (2025)
41. Zhang, Z., Liao, C., Zhang, H., Chen, H.H., Chen, K., Wen, Z., Guo, L., Ren, B., Zheng, X., Li, Y., et al.: Panoramic affordance prediction. arXiv preprint arXiv:2603.15558 (2026)
42. Zheng, C., Zhang, Z., Zhang, B., Lin, R., Lu, K., Yu, B., Liu, D., Zhou, J., Lin, J.: Processbench: Identifying process errors in mathematical reasoning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1009–1024 (2025)
43. Zheng, X., Dongfang, Z., Jiang, L., Zheng, B., Guo, Y., Zhang, Z., Albanese, G., Yang, R., Ma, M., Zhang, Z., et al.: Multimodal spatial reasoning in the large model era: A survey and benchmarks. arXiv preprint arXiv:2510.25760 (2025)
44. Zheng, X., Liao, C., Weng, Z., Lei, K., Dongfang, Z., He, H., Lyu, Y., Jiang, L., Qi, L., Chen, L., et al.: Panorama: The rise of omnidirectional vision in the embodied ai era. arXiv preprint arXiv:2509.12989 (2025)
45. Zhong, D., Zheng, X., Liao, C., Lyu, Y., Chen, J., Wu, S., Zhang, L., Hu, X.: Omnisam: Omnidirectional segment anything model for uda in panoramic semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23892–23901 (2025)
46. Zhou, Y., Zhang, T., Zhang, D., Ji, S., Li, X., Qi, L.: Dense360: Dense understanding from omnidirectional panoramas. arXiv preprint arXiv:2506.14471 (2025)