SEERBench: A Spatial Ego-Exo Reasoning Benchmark for MLLMs

Fengyuan Lu[1*], Jiahe Feng[2*], Zhengyang Zhou[1*], Shaofeng Zhang[3]
Wenbin Li[1], Qi Fan[1✉], and Yang Gao[1]

[1]Nanjing University [2]Tianjin University [3]University of Science and Technology of China
fanqi@nju.edu.cn

Abstract. Humans seamlessly integrate egocentric perceptions with exocentric perspectives to comprehend dynamic spatial positioning. For Embodied AI, mastering this ego-exo spatial reasoning is a fundamental prerequisite for real-world interaction. Although Multimodal Large Language Models (MLLMs) have demonstrated significant progress in general visual reasoning, their capacity for ego-exo spatial reasoning in real-world environments remains largely unexplored. To investigate whether current MLLMs exhibit such human-like spatial cognition, we introduce **SEERBench**, a rigorous benchmark comprising 1,198 high-quality, human-annotated question-answer pairs across 35 diverse real-world Ego-Exo scenarios. SEERBench evaluates models through 10 distinct tasks structured into three progressive cognitive levels: Spatial Perception, Spatial Imagination, and Ego-Exo Collaboration. Extensive evaluations of 19 representative MLLMs reveal a substantial performance gap: even the best-performing frontier models trail human experts by a margin of 46.8%, highlighting severe limitations in current multimodal foundations. To address this, we propose **SEER-Map**, a training-free, tool-augmented baseline. By explicitly constructing top-down Bird's-Eye View (BEV) spatial topologies to assist in ego-exo geometric alignment, SEER-Map functions as an Oracle-informed probing baseline lower the threshold of spatial understanding. Codes and data are available here.

Keywords: Multimodal Large Language Models · Ego-Exo Spatial Reasoning · Vision-Language Benchmark

1 Introduction

Humans possess an innate capacity for ego-exo spatial reasoning, seamlessly integrating egocentric perceptions with exocentric perspectives to transcend the perceptual limitations of singular viewpoints [6,39]. By anchoring local observations within broader environmental structures, this synergy provides a robust cognitive foundation for complex spatial deduction and purposeful interaction [45]. Replicating this capability is pivotal for Embodied AI, empowering autonomous agents to comprehend dynamic spatial positioning and facilitating multi-agent collaboration through perspective-sharing [32, 44].

*Equal contribution. ✉Corresponding author.

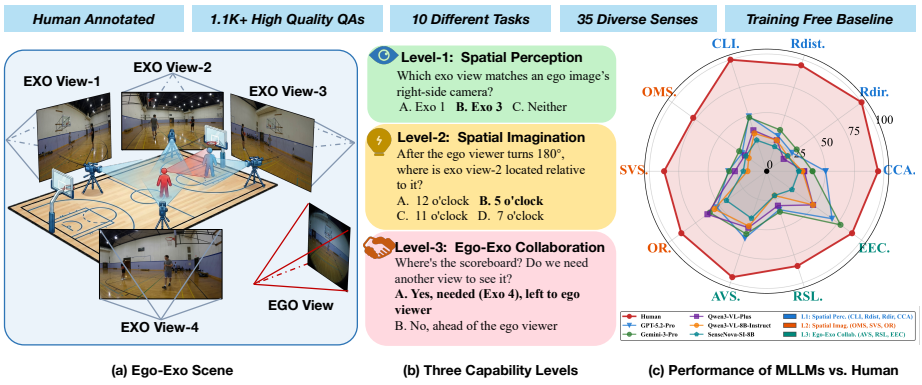

Fig. 1: Illustration of the SEER Benchmark. **Left**: A basketball court scene depicting one ego-centric and four exo-centric camera views. **Middle**: Examples of QA pairs designed to assess three capability levels in SEERBench. **Right**: Accuracy comparison between MLLMs and humans on SEERBench.

Despite the transformative breakthroughs of Multimodal Large Language Models (MLLMs) in general visual reasoning and action planning [1, 4, 8, 11, 26], whether these models possess human-like ego-exo spatial intelligence remains an open question. Specifically, while current MLLMs excel in semantic recognition, they struggle significantly to bridge the geometric gap between disparate viewpoints, a fundamental deficiency that hinders the pursuit of advanced spatial cognition [41, 47].

In the landscape of multimodal evaluation, the critical intersection of spatial reasoning and ego-exo perspective integration remains largely unexplored. On one hand, prominent spatial benchmarks [37, 40, 47] typically probe reasoning within isolated, single-view frames, overlooking the rigorous geometric challenges inherent in ego-exo view alignment. On the other hand, existing ego-exo initiatives [16, 22] focus predominantly on semantic-level tasks like action recognition, failing to evaluate joint spatial deduction across views. Consequently, a systematic framework to investigate this untapped intersection, where models must resolve spatial ambiguities through active ego-exo collaboration, is conspicuously missing.

To bridge this gap, we introduce SEERBench, a specialized benchmark for evaluating ego-exo spatial reasoning. Developed through manual scene screening and human-led logical verification , the benchmark comprises 1,198 high-quality, human-annotated question-answer pairs across 35 diverse real-world scenarios. We propose a hierarchical evaluation framework structured into three progressive cognitive levels: Spatial Perception, Spatial Imagination, and Ego-Exo Collaboration, encompassing 10 tasks. While prior benchmarks focus on isolated or symmetric frames , our Collaboration level fills a critical gap by simulating multi-agent scenarios. It requires MLLMs to transcend individual observational limits to achieve synergistic spatial consensus, a prerequisite for autonomous coordina-

tion in complex environments. To verify these capabilities, we comprehensively evaluate 19 representative MLLMs, including state-of-the-art closed-source and open-source models, against human performance. As illustrated in Fig. 1, extensive evaluations reveal a substantial performance gap: while human experts achieve 91.5% average accuracy, the best-performing model reaches only 44.7%, highlighting the persistent challenges MLLMs face in achieving robust ego-exo spatial intelligence.

To investigate potential solutions for this severe performance gap, we further propose SEER-Map, a training-free, tool-augmented baseline. By leveraging 3D vision foundation models [43] to extract global geometry and render a top-down Bird's-Eye View (BEV) spatial topology, SEER-Map explicitly provides MLLMs with a unified spatial prior. Experimental results demonstrate that this explicit geometric grounding yields substantial and consistent performance gains across various model families.

In summary, our main contributions are threefold:

- We introduce **SEERBench**, a benchmark specifically designed to evaluate ego-exo spatial reasoning in MLLMs, featuring a hierarchical taxonomy that ranges from basic spatial perception to complex multi-view collaboration.
- We evaluate 19 representative MLLMs on SEERBench. The results reveal a substantial performance gap between current models and human experts, with the best-performing MLLMs trailing human performance by an absolute margin of 46.8%.
- We propose **SEER-Map**, a training-free and Oracle-informed baseline that lowers the spatial perception threshold, yielding consistent performance gains across diverse MLLMs.

2 Related Work

Ego-Exo Data and Benchmarks. While datasets like Ego-Exo4D [15] established the foundation for joint understanding, recent benchmarks [17, 19, 21, 38] predominantly evaluate semantic-level comprehension (e.g., action recognition or temporal consistency). As summarized in Table 1, these tasks can often be bypassed via 2D semantic matching. In contrast, SEERBench mandates rigorous 3D spatial alignment and mental viewpoint simulation, shifting the paradigm towards explicit cross-view geometric deduction.

Spatial Understanding Benchmarks and Models. Existing spatial benchmarks typically focus on single-view observations [3, 10, 12, 20, 25, 27, 40, 47] (Table 1). Even multi-camera extensions like MMSI-Bench [48] assume symmetric setups, neglecting the complex, asymmetric geometries inherent in ego-exo human interactions. On the modeling front, although specialized MLLMs enhance reasoning via explicit coordinates [9], grounded pre-training [7], or reinforcement learning [29], these gains remain confined to single-view perception. They persistently struggle to establish accurate geometric correspondences across unaligned ego-exo viewpoints, highlighting the critical need for diagnostic frameworks like SEERBench.

Table 1: Comparison of existing Ego-Exo and Spatial Understanding benchmarks. Unlike prior datasets bounded by semantic recognition or single-view constraints, SEERBench explicitly shifts the paradigm to rigorous, multi-view geometric deduction across three cognitive levels.

Benchmark	Ego-Exo Setting	Primary Focus	Spatial Perc.	Spatial Imag.	EgoExo Collab.
ScanQA [3]	✗	3D Scene QA	✓	✗	✗
SpatialMQA [27]	✗	Single-view Spatial	✓	✗	✗
CA-VQA [10]	✗	3D Scene QA	✓	✓	✗
MMSI-Bench [48]	✗	Multi-view Spatial	✓	✓	✗
EgoExoBench [17]	✓	Action & Semantics	✓	✗	✗
E3VQA [21]	✓	General VQA	✓	✗	✗
EgoExo-Con [19]	✓	Temporal Consistency	✓	✗	✗
Ego-2-Exo [38]	✓	Ego-property Transfer	✓	✗	✗
SEERBench (Ours)	✓	**Ego-Exo Spatial Reasoning**	✓	✓	✓

Tool-augmented Spatial Reasoning. Test-time tool augmentation enhances MLLM spatial reasoning via explicit priors without retraining [49]. Existing systems primarily leverage three tool categories: (i) UI-like visual primitives for extracting fine-grained cues [50]; (ii) 2D perception modules that generate object-centric facts or BEV sketches [51]; and (iii) 3D view synthesis for viewpoint-controllable evidence generation [28, 31, 34]. Recently, agentic controllers have automated tool execution via policy planning [33, 46], revealing that metrically grounded feedback is critical for reasoning across non-co-visible frames [36]. Building on this, we propose SEER-Map, a *training-free* DUSt3R-based [43] geometric stack. By projecting complex scene structures into a minimalist BEV topology map, SEER-Map explicitly reduces the geometric reasoning burden for MLLMs during ego-exo visual question answering.

3 SEERBench

3.1 Overview

Ego-exo spatial reasoning requires integrating first-person perceptions with third-person contexts to comprehend 3D environments. Evaluating this spatial cognition is the primary objective of **SEERBench**.

SEERBench comprises 1,198 human-annotated question-answer (QA) pairs across 35 real-world scenes and 10 distinct tasks. For standardized evaluation, questions are formulated as multiple-choice with 3–4 options, including both single- and multi-choice formats. This dataset design enables the systematic assessment of Multimodal Large Language Models (MLLMs) across various physical activities and environmental layouts.

The 10 evaluation tasks are structured into three progressive cognitive levels: Spatial Perception, Spatial Imagination, and Ego-Exo Collaboration. This hierarchical taxonomy evaluates capabilities ranging from basic geometric grounding to multi-view information aggregation. Finally, we establish a benchmark baseline by evaluating 19 closed-source and open-source MLLMs.

Fig. 2: Overview of SEERBench. A comprehensive benchmark for ego-exo spatial reasoning, spanning three hierarchical levels and ten distinct tasks. Correct answers are indicated in red.

3.2 Hierarchical Taxonomy

To provide a structured evaluation of MLLMs in multi-view, ego-exo environments, SEERBench introduces a taxonomy comprising three progressive cognitive levels. This framework assesses spatial reasoning from foundational geometric grounding to complex multi-view aggregation.

Level 1: Spatial Perception. This level assesses MLLMs' basic capability to extract and ground 3D geometric relationships from static ego-exo observations. It evaluates self-localization, spatial orientation recognition, and exocentric camera-to-content mapping. This foundational perception is essential for embodied agents to comprehend surroundings and support tasks like obstacle avoidance. Key proficiencies include *Camera-Content Alignment*, *Relative Distance Estimation*, *Relative Direction Estimation*, and *Camera Layout Inference*.

Level 2: Spatial Imagination. Advancing beyond static observation, this level challenges MLLMs to synthesize ego-exo information to simulate dynamic changes—such as hypothetical viewpoint shifts or object movements—and predict subsequent environmental states. In real-world applications, this predictive capacity is crucial for motion planning, allowing agents to anticipate the geometric consequences of their actions without relying on costly physical trial-and-error. Evaluated tasks include *Object Motion Simulation*, *Simulated Viewpoint Shift*, and *Occlusion Reasoning*.

Level 3: Ego-Exo Collaboration. This level evaluates the capacity of MLLMs to aggregate and reason over distributed visual information. It probes whether models can identify informative cross-view correlations to resolve spatial ambiguities or conflicting visual cues. The practical value lies in handling complex environments where severe occlusions or single-view limitations make individual perspectives unreliable. Tasks in this tier include *Active Viewpoint Selection*, *Ego-Exo Consistency*, and *Robust Spatial Logic*.

3.3 Data Construction Pipeline

(1) Data Sourcing and Curation. We construct strictly spatiotemporally synchronized ego-exo frame tuples, drawing exclusively from the validation and test sets of EgoExo4D. These segments underwent exhaustive manual screening to isolate informative interactions, followed by a filtering step to remove occlusions, motion blur, and uninformative content, thereby ensuring a high-fidelity foundation for ego-exo spatial reasoning.

(2) Hybrid Fine-grained Annotation. Curated frames undergo pixel-level annotation via a collaborative AI-human approach. AI tools accelerate spatial metadata extraction, followed by manual refinement to ensure precision. The resulting metadata includes fine-grained keypoints, 2D bounding boxes, and pixel-accurate masks for target objects.

(3) Dual-Track QA Generation. QA pairs are produced through two distinct workflows: an *AI-generative track* that leverages LLMs to produce diverse candidates from annotated facts, and an *expert-template track* where human specialists design rigorous logical templates for intricate geometric reasoning and perspective-shifting scenarios.

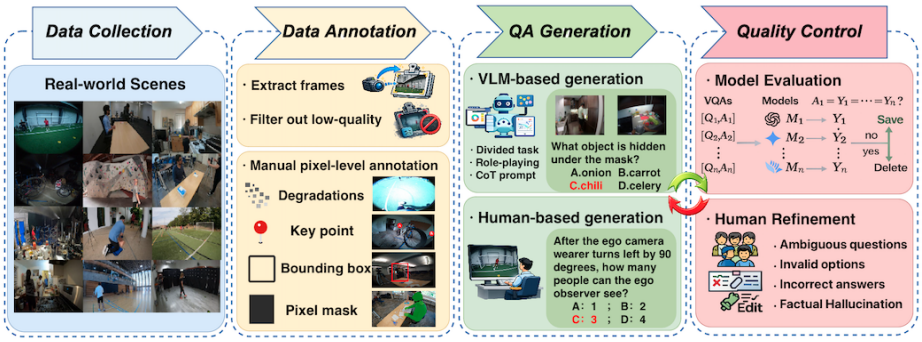

Fig. 3: The SEERBench construction pipeline. Our workflow integrates automated AI metadata extraction with expert human refinement across four key phases: data curation, fine-grained pixel annotation, dual-track QA generation, and closed-loop quality control. The circular arrows signify our iterative refinement process to ensure the benchmark's rigor and difficulty.

(4) Multi-Dimensional Quality Control. To maintain benchmark difficulty, we implement modality-specific filtering to discard "shortcut" questions solvable via single-view or text-only cues. After joint AI and human verification, low-quality samples are fed back to the annotation stage, creating a closed-loop refinement to iteratively optimize the entire pipeline.

3.4 Benchmark Statistics

Table 2 summarizes SEERBench's statistics, comprising 1,198 human-annotated QA pairs and 3,721 unique images across 35 real-world scenes. To support rigorous evaluation, questions average 105.9 words and leverage multiple multi-view images. Figure 4 illustrates the data distribution across our cognitive taxonomy: Spatial Perception (49.7%), Spatial Imagination (35.9%), and Ego-Exo Collaboration (14.4%). This hierarchical distribution ensures comprehensive coverage of progressive spatial reasoning capabilities. We further expand the Level-3 subset and report 95% bootstrap confidence intervals in Supplementary Material.

4 Methodology: SEER-Map Baseline

Despite their proficiency in single-view perception, current MLLMs often struggle to maintain a coherent global understanding in multi-view, ego-exo environments. This limitation stems from the difficulty of establishing cross-view correspondences within implicit pixel spaces. To address this, we introduce **SEER-Map**, a training-free, tool-augmented baseline that provides MLLMs with an explicit "God-view" spatial prior. By transforming complex multi-view physical geometry into an intuitive 2D bird's-eye view (BEV) map, SEER-Map reduces the geometric reasoning burden for the model.

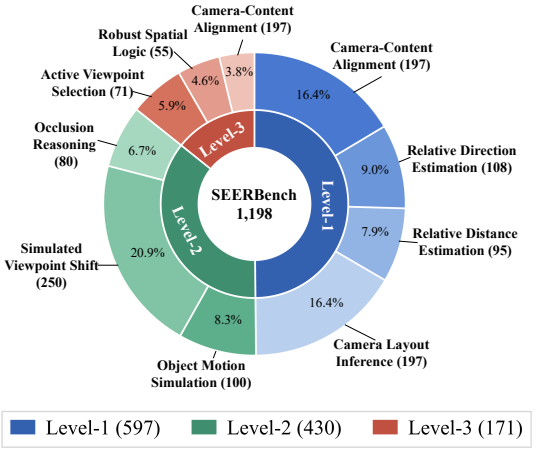

Fig. 4: Distribution of SEERBench QA pairs across tasks and cognitive levels.

Table 2: Summary statistics of SEERBench. We report the number of QA pairs, images, annotated images, scene count, and per-question image/text statistics.

Statistic	Value
Total QAs	1,198
Unique Imgs	3,721
Annotated Imgs	726
Avg. QA Length	105.9
Avg. Imgs / QA	3.4
Max Imgs / QA	6
Unique Scenes	35

4.1 Spatial Topology Generation

Given a synchronized set $\mathcal{I} = \{I_{ego}, I^1_{exo}, \ldots, I^N_{exo}\}$ of ego- and exo-centric views and the 2D pixel coordinates of a target entity (e.g., an object or human, where specified by the task), our mapping pipeline generates a 2D spatial topology map $\mathcal{M}_{topo}$ through three coherent steps:

1. Global Geometry Extraction. We employ DUSt3R as a representative 3D vision foundation model to recover the scene's geometry without requiring camera intrinsics. The model processes the image set $\mathcal{I}$ to recover the relative camera extrinsics and a dense 3D point cloud within a unified world coordinate system. This step establishes the foundational geometric alignment necessary for subsequent cross-view reasoning.

2. 3D Grounding of Target Entities. For tasks requiring target localization, the 2D pixel coordinate in the reference view is utilized to index into the corresponding 3D point cloud. Through this lookup operation, the 2D pixel is mapped precisely to an absolute 3D physical coordinate in the real world, effectively grounding semantic entities into the physical environment.

3. Topological Projection and Rendering. After 3D spatial information is acquired, the height dimension is discarded to simplify the scene. The 2D camera frustums are computed from extrinsics, and the 3D coordinate of the target (if available) is projected orthogonally onto the ground plane. Finally, these 2D elements are rendered on a blank canvas, generating the minimalist topology map $\mathcal{M}_{topo}$ that encapsulates the relative poses of all cameras and target.

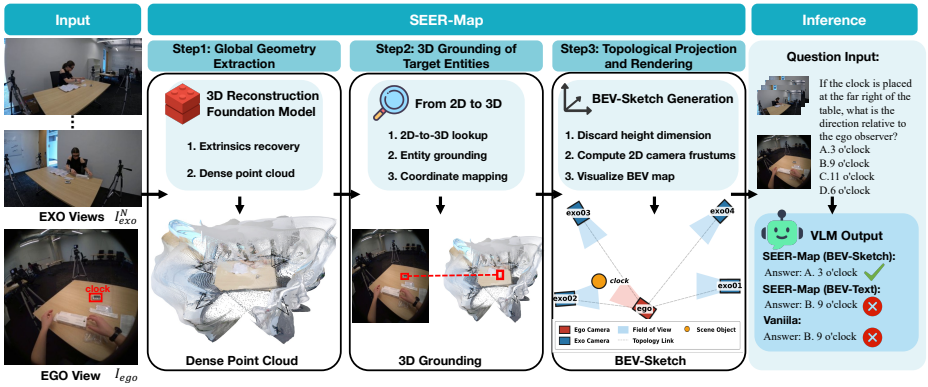

Fig. 5: Overview of the SEER-Map pipeline. Given a synchronized set of ego-exo views, SEER-Map first extracts global scene geometry to generate a dense point cloud. Next, 2D target entities are grounded into 3D physical coordinates. Finally, the 3D spatial information is orthogonally projected to render a minimalist 2D Bird's-Eye View (BEV) topology map. This top-down BEV sketch is then provided to the VLM as an explicit spatial prior during inference, significantly reducing the geometric reasoning burden.

4.2 Augmented Visual Prompting

In practice, the generated SEER-Map $\mathcal{M}_{topo}$, which contains the global spatial prior, is integrated with the original multi-view image set $\mathcal{I}$ to serve as interleaved visual inputs for the MLLM. When paired with specific text instructions, this explicit top-down perspective enables the model to bypass the low-level challenges of geometric alignment and to focus directly on higher-order spatial reasoning and multi-view collaboration tasks. By providing this intuitive "God-view" prior, SEER-Map effectively transforms the complex cross-view association burden into a simplified topological matching problem. This training-free augmentation ensures that the model can leverage its inherent semantic reasoning strengths while operating within a metrically grounded spatial context.

5 Experimental Results

5.1 Evaluation Setups

Evaluated Models. We benchmark MLLMs on SEERBench across three categories: **(i) closed-source models**, including GPT-4o [18], GPT-5.2 (and GPT-5.2-Pro) [35], Gemini-3 (Flash/Pro) [13,14], Claude-Sonnet-4.5 [2], and Qwen3-VL-Plus [5]; **(ii) open-source models**, including Qwen3-VL [5], DeepSeek-VL [30], InternVL3 [52], GLM-4.6V [42], and LLaVA variants [23,24]; and **(iii) spatial-oriented models**, including SenseNova-SI and Spatial-SSRL [7,29]. All closed-source models are evaluated via their official APIs. All open-source and spatial-oriented models are evaluated on 8 NVIDIA RTX 4090 GPUs, strictly

following the official inference configurations released by the corresponding authors/model cards. For more experimental details, please refer to the Supplementary Material.

Human Evaluation. We evaluate human performance on a randomly selected subset of 200 instances, comprising 20 samples for each capability sub-task.

Task Abbreviations. Table 3 reports results using compact task abbreviations. **Spatial Perception** includes *Camera-Content Alignment* (***CCA.***), *Relative Direction Estimation* (***RDir.***), *Relative Distance Estimation* (***RDist.***), and *Camera Layout Inference* (***CLI.***). **Spatial Imagination** includes *Object Motion Simulation* (***OMS.***), *Simulated Viewpoint Shift* (***SVS.***), and *Occlusion Reasoning* (***OR.***). **Ego-Exo Collaboration** includes *Active Viewpoint Selection* (***AVS.***), *Robust Spatial Logic* (***RSL.***), and *Ego-Exo Consistency* (***EEC.***).

5.2 Main Results

Table 3 summarizes the main comparison across all evaluated models.

Current MLLMs remain far below human performance. Despite strong progress in generic multimodal reasoning, all tested model families underperform human experts by a wide margin on SEERBench. Human experts reach **91.5**% average accuracy, while the best-performing closed-source model Gemini-3-Pro achieves only **44.7**% Avg., leaving an absolute gap of **46.8** points. Moreover, several model–task pairs fall close to chance-level performance (roughly ∼25–33%), highlighting that ego–exo spatial reasoning remains a frontier challenge rather than a solved capability.

Comparison across models. We observe a consistent ranking across the three model categories. Closed-source models lead overall, with Gemini-3-Pro achieving the best average accuracy of 44.7, and GPT-5.2-Pro following closely with 43.3. The strongest open-source model, Qwen3-VL-32B-Thinking, reaches 36.6, indicating a clear gap to the top proprietary systems while remaining competitive among open models. In comparison, spatial-oriented models lag behind general-purpose MLLMs, as SenseNova-SI-8B obtains 27.5 and Spatial-SSRL-4B further drops to 22.2, suggesting that current spatial specialization does not yet translate into strong ego-exo reasoning on SEERBench.

Task-wise behaviors. We identify consistent behavioral patterns across all evaluated model families:

(1) Spatial perception tasks can be partially solved but remain fragile. Strong closed-source models achieve the highest Level-1 scores on camera-layout inference, with GPT-5.2-Pro reaching 49.2. Yet most models perform poorly on relative direction and distance estimation, indicating difficulty in establishing precise metric or egocentric geometric grounding from paired views.

(2) Spatial imagination serves as the primary bottleneck. Object-motion simulation and simulated viewpoint shift bring universal challenges. Even top closed-source models only achieve low-30s accuracy, while many open-source models fall below 20 percent. This reveals limited ability for mental viewpoint transformation and state transition prediction under spatial constraints.

Table 3: Accuracy of evaluated MLLMs on SEERBench. For each capability category, we report the overall accuracy. The best result within each category is marked in bold and the second-best is underlined.

Method	Spatial Perc.				Spatial Imag.			EgoExo Collab.			Avg.	Rank
	CCA.	RDir.	RDist.	CLI.	OMS.	SVS.	OR.	AVS.	RSL.	EEC.		
Human Performance												
Human Expert	95.0	100.0	95.0	100.0	77.5	87.5	90.0	95.0	85.0	90.0	91.5	-
Closed-source Models												
GPT-4o	19.8	26.3	30.6	38.6	**34.0**	22.4	56.3	40.8	27.3	33.3	32.9	6
GPT-5.2	41.6	17.9	25.9	39.1	26.0	26.8	52.5	**63.4**	32.7	55.6	38.2	4
GPT-5.2-Pro	**50.3**	26.7	31.6	**49.2**	26.0	**32.8**	52.5	60.1	34.5	68.9	43.3	2
Gemini-3-Flash	32.0	26.3	35.2	42.6	28.0	30.4	58.8	60.6	32.7	75.6	42.2	3
Gemini-3-Pro	39.1	**31.6**	**37.0**	48.2	29.0	32.0	60.0	56.3	**36.4**	**77.8**	**44.7**	1
Claude-Sonnet-4.5	26.4	15.8	23.1	48.2	23.0	18.8	52.5	46.5	25.5	44.4	32.4	7
Qwen3-VL-Plus	32.0	17.9	27.8	36.7	23.0	27.2	**62.5**	50.7	30.9	48.9	35.8	5
Open-source Models												
Qwen3-VL-8B-Instruct	30.5	22.1	27.8	34.0	19.0	16.4	55.0	49.3	21.8	48.9	32.5	4
Qwen3-VL-8B-Thinking	26.4	25.4	27.4	25.4	12.0	17.2	25.0	35.2	27.3	20.0	24.1	9
Qwen3-VL-32B-Instruct	31.5	26.3	22.2	33.5	23.0	**24.8**	63.8	52.1	29.1	53.3	36.0	2
Qwen3-VL-32B-Thinking	27.9	21.1	28.7	33.5	18.0	18.4	**65.0**	**63.4**	27.3	**62.2**	**36.6**	1
DeepSeek-VL-7B-Chat	32.5	23.2	18.9	31.0	16.0	15.2	46.3	22.5	16.4	37.8	26.0	7
GLM-4.6V	27.9	28.7	25.3	**39.1**	**28.0**	18.0	56.3	49.3	**32.7**	48.9	35.4	3
LLaVA-OneVision-Qwen2-7B	28.4	17.9	29.6	38.1	14.0	12.8	45.0	38.0	7.3	26.7	25.0	8
LLaVA-NeXT-7B	26.4	11.6	**33.3**	33.0	11.0	10.8	43.8	31.0	9.1	22.2	23.2	10
InternVL3-8B	**34.5**	**30.5**	18.5	33.0	16.0	11.6	42.5	28.2	12.7	33.3	26.1	6
InternVL3-14B	32.0	**30.5**	20.4	33.0	19.0	15.2	41.3	35.2	20.0	26.7	27.3	5
Spatial Understanding Models												
SenseNova-SI-8B	**27.4**	**22.2**	22.1	**27.9**	**22.0**	**20.0**	**42.5**	**42.2**	**21.8**	**26.7**	**27.5**	1
Spatial-SSRL-4B	17.3	17.6	**25.3**	25.4	20.0	13.6	37.5	26.8	16.4	22.2	22.2	2

(3) Ego–Exo collaboration shows a clear performance split. Models perform relatively well on ego–exo consistency and active viewpoint selection. Gemini-3-Pro attains 77.8 on the former, and GPT-5.2 reaches 63.4 on the latter. By contrast, robust spatial logic remains highly difficult. Many open-source models collapse to single-digit or low-teens accuracy, with LLaVA-OneVision-Qwen2-7B scoring only 7.3. This suggests failures in multi-hop geometric deduction beyond basic recognition or heuristic cue matching.

5.3 Performance Evaluation of SEER-Map

As shown in Table 4, SEER-Map yields substantial, consistent gains across closed-source, open-source, and spatial-oriented models. Average accuracy in-

Table 4: Effect of SEER-Map on SEERBench. We report accuracy before and after applying SEER-Map for closed-source, open-source, and spatial-oriented models, demonstrating consistent gains from the top-down topology map augmentation.

Method	Spatial Perc.				Spatial Imag.			EgoExo Collab.			Avg.
	CCA.	RDir.	RDist.	CLI.	OMS.	SVS.	OR.	AVS.	RSL.	EEC.	
GPT-5.2	41.6	17.9	25.9	39.1	26.0	26.8	52.5	63.4	32.7	55.6	38.2
+ SEER-Map	54.8	29.6	43.2	82.2	36.0	37.2	71.3	66.2	32.7	57.8	51.1
Δ	+13.2	+11.7	+17.3	+43.1	+10.0	+10.4	+18.8	+2.8	+0.0	+2.2	+13.1
Gemini-3-Flash	32.0	26.3	35.2	42.6	28.0	30.4	58.8	60.6	32.7	75.6	42.2
+ SEER-Map	56.3	36.1	41.0	81.2	32.0	38.0	68.8	59.2	34.5	75.6	52.8
Δ	+24.3	+9.8	+5.8	+38.6	+4.0	+7.6	+10.0	−1.4	+1.8	+0.0	+10.6
Qwen3-VL-Plus	32.0	17.9	27.8	36.7	23.0	27.2	62.5	50.7	30.9	48.9	35.8
+ SEER-Map	40.1	24.1	31.6	75.0	27.0	40.8	68.8	54.9	27.3	48.9	43.9
Δ	+8.1	+6.2	+3.8	+38.3	+4.0	+13.6	+6.3	+4.2	−3.6	+0.0	+8.1
Qwen3-VL-8B-Instruct	30.5	22.1	27.8	34.0	19.0	16.4	55.0	49.3	21.8	48.9	32.5
+ SEER-Map	37.1	25.9	31.6	73.1	25.0	32.8	60.0	56.3	12.7	46.7	40.1
Δ	+6.6	+3.8	+3.8	+39.1	+6.0	+16.4	+5.0	+7.0	−9.1	−2.2	+7.6
SenseNova-SI-8B	27.4	22.2	22.1	27.9	22.0	20.0	42.5	42.2	21.8	26.7	27.5
+ SEER-Map	30.5	24.1	29.5	76.1	26.0	25.6	55.0	42.2	25.5	26.7	36.1
Δ	+3.1	+1.9	+7.4	+48.2	+4.0	+5.6	+12.5	+0.0	+3.7	+0.0	+8.6

creases from 38.2 to 51.1 for GPT-5.2 and from 42.2 to 52.8 for Gemini-3-Flash. Similar improvements on Qwen3-VL-Plus, Qwen3-VL-8B-Instruct, and SenseNova-SI-8B demonstrate that these benefits generalize across model families. Notably, SEER-Map most strongly boosts camera-layout inference, providing a reliable global topology cue for ego–exo alignment.

5.4 Diagnosing VLM Limitations and the Efficacy of SEER-Map

The primary failure of MLLMs in ego-exo scenarios is a cross-view geometric disconnect. While recognizing semantics in individual frames, they fail to associate exocentric observations with the egocentric observer's position. Consequently, MLLMs process unaligned views in isolation, leading to flawed spatial hallucinations (Fig. 7). SEER-Map overcomes this bottleneck by decoupling perception from reasoning. Instead of forcing MLLMs to implicitly infer alignments, it delegates 3D reconstruction to specialized vision models. Providing an explicit Bird's-Eye View (BEV) topology eliminates low-level geometric burdens, allowing MLLMs to bypass alignment hurdles and focus entirely on higher-order spatial reasoning.

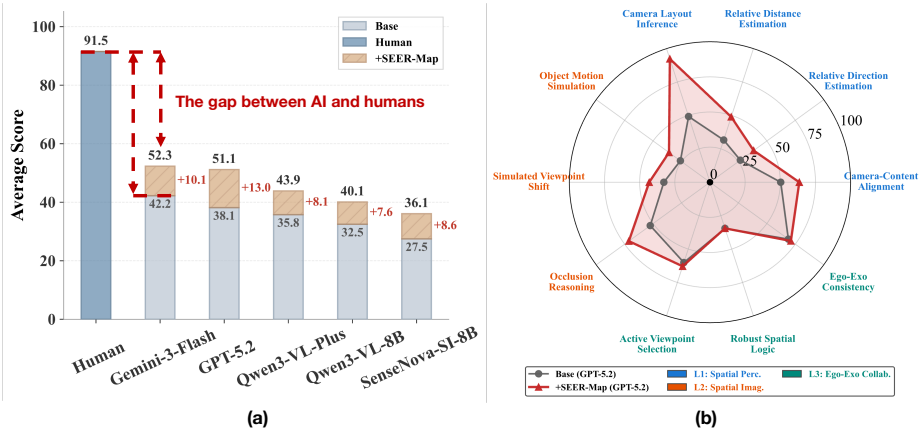

Fig. 6: (a) **Overall Comparison**: Average accuracy of MLLMs with/without SEER-Map vs. humans. SEER-Map yields consistent gains despite a persistent human-AI gap. (b) **Task-wise Breakdown**: GPT-5.2 performance across 10 sub-tasks. Explicit spatial priors significantly boost geometric-heavy tasks like Camera Layout Inference.

Table 5: Comparison of Information-Equivalent Modalities. We compare text-based 3D coordinates against SEER-Map's visual BEV sketches. Despite conveying identical geometric information, explicit visual representations significantly reduce cross-modal cognitive load and improve reasoning accuracy.

Model	Target Grounding	Global Spatial Prior
	Text → **Visual (Ours)** (Δ)	Text → **Visual BEV (Ours)** (Δ)
Gemini-3-Flash	40.1 → **44.2** (+4.1)	44.6 → **52.3** (+7.7)
Qwen3-VL-8B-Instruct	27.4 → **32.5** (+5.1)	33.1 → **40.1** (+7.0)
SenseNova-SI-8B	23.7 → **27.5** (+3.8)	28.4 → **36.1** (+7.7)

5.5 Visual vs. Textual Spatial Priors: An Information-Equivalent Comparison

To ensure a fair evaluation of our visual formulation, we introduce an *information-equivalent text baseline* by replacing the visual BEV map with its exact numerical 3D geometry (i.e., camera extrinsics and 3D target coordinates). As shown in Table 5, despite conveying identical mathematical geometry, explicit visual representations significantly outperform numerical text (e.g., a **+7.7%** absolute boost for Gemini-3-Flash). This exposes a severe cross-modal cognitive bottleneck: mentally reconstructing 3D topologies from raw text overwhelms current MLLMs. By directly rendering this geometry as a 2D sketch, SEER-Map effectively decouples low-level spatial grounding from high-order geometric reasoning.

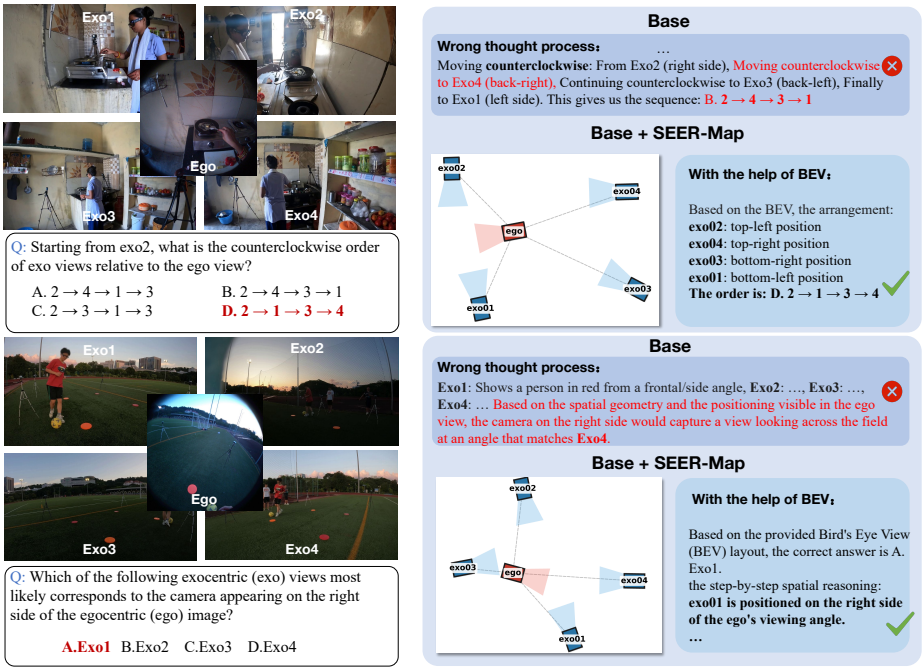

Fig. 7: Qualitative comparison between standard MLLMs and SEER-Map . While base models hallucinate spatial relationships across unaligned views , SEER-Map's BEV topology anchors the global layout to resolve geometric ambiguities.

5.6 Failure-Mode Analysis

We analyze 100 randomly sampled Gemini-3-Pro failures to identify bottlenecks on SEERBench. As shown in Table 6, cross-view mismatch is the most frequent error (31%), followed by reasoning breakdown (29%), indicating that MLLMs primarily struggle with ego-exo correspondence and multi-step spatial deduction rather than pure perception or hallucination.

Further layer-wise attention analysis on InternVL3-4B (Fig. 8) reveals that incorrect predictions often over-attend to misleading exocentric views while underattending to egocentric or ground-truth supporting views. This confirms that SEERBench failures stem from poor cross-view alignment and attention allocation, highlighting the necessity of SEER-Map's explicit BEV topology as a global spatial prior.

6 Conclusion

In this work, we present **SEERBench**, a comprehensive benchmark designed to evaluate ego-exo spatial reasoning in MLLMs. Comprising 1,198 human-annotated QA pairs across 10 tasks and three cognitive levels, it systematically

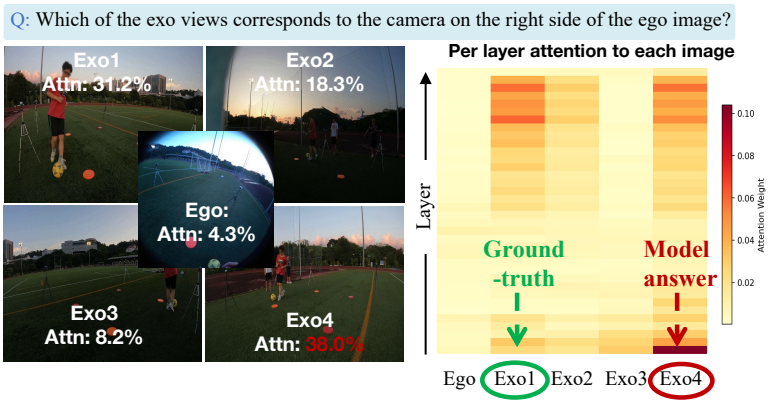

Fig. 8: Layer-wise visual attention of InternVL3-4B on a failure case. The model over-attends to misleading exocentric views but under-attends to egocentric and ground-truth supporting views, revealing misallocated attention in ego-exo reasoning.

Table 6: Failure-mode analysis on 100 randomly sampled Gemini-3-Pro errors.

Failure Mode	Description	Ratio
Cross-view mismatch	Incorrect object mapping across views	31%
Reasoning breakdown	Logic errors in reasoning	29%
Spatial confusion	Misjudged positions or directions	22%
Perception error	Misidentified target objects	10%
Spatial hallucination	Reasoning without evidence	8%

diagnoses cross-view spatial intelligence. Our evaluation of 19 representative MLLMs reveals a substantial 46.8% absolute performance gap behind human experts, exposing a critical geometric deficiency in current multimodal foundations. To bridge this gap, we propose **SEER-Map**, a training-free baseline that explicitly constructs top-down BEV spatial topologies for ego-exo alignment, demonstrating consistent reasoning improvements across diverse model families. We hope SEERBench catalyzes the development of robust, spatially-aware foundation models for Embodied AI.

7 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China, under Grant Nos. 62192783, 62276128, and 62406140; the Young Elite Scientists Sponsorship Program by China Association for Science and Technology, under Grant No. 2023QNRC001; the Key Research and Development Program of Jiangsu Province, under Grant No. BE2023019; and the Jiangsu Natural Science Foundation, under Grant Nos. BK20221441 and BK20241200. The authors would like to thank the support of Huawei Ascend Cloud Ecological Development Project.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 2
2. Anthropic: Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5 (2025), accessed: 2026-03-05 9
3. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19107–19117 (2021), https://api.semanticscholar.org/CorpusID:245334889 3, 4
4. 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) 2
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 9
6. Burgess, N.: Spatial memory: how egocentric and allocentric combine. Trends in cognitive sciences **10**(12), 551–557 (2006) 1
7. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., Zhou, T., Li, J., Pang, H.E., Qian, O., Wei, Y.R., Lin, Z., Shi, X., Deng, K., Han, X., Chen, Z., Fan, X., Deng, H., Lu, L., Pan, L., Li, B., Liu, Z., Wang, Q., Lin, D., Yang, L.: Scaling spatial intelligence with multimodal foundation models. ArXiv **abs/2511.13719** (2025), https://api.semanticscholar.org/CorpusID:283073248 3, 9
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: 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. pp. 24185–24198 (2024) 2
9. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision language model. ArXiv **abs/2406.01584** (2024), https://api.semanticscholar.org/CorpusID:270215984 3
10. 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) 3, 4
11. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023) 2
12. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. ArXiv **abs/2404.12390** (2024), https://api.semanticscholar.org/CorpusID:269214091 3
13. Google Gemini Team: Gemini 3 flash model card. Technical report, Google DeepMind (2026), https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf 9
14. Google Gemini Team: Gemini 3 pro model card. Technical report, Google DeepMind (2026), https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf 9

15. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., et al.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19383–19400 (2023), https://api.semanticscholar.org/CorpusID:265506384 3
16. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first- and third-person view video understanding in mllms. ArXiv **abs/2507.18342** (2025), https://api.semanticscholar.org/CorpusID:280017247 2
17. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first- and third-person view video understanding in mllms. arXiv preprint arXiv:2507.18342 (2025) 3, 4
18. 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 (2024) 9
19. Jung, M., Xiao, J., Kim, J., Zhang, B.T., Yao, A.: Egoexo-con: Exploring view-invariant video temporal understanding. arXiv preprint arXiv:2510.26113 (2025) 3, 4
20. Kamath, A., Hessel, J., Chang, K.W.: What's "up" with vision-language models? investigating their struggle with spatial reasoning. ArXiv **abs/2310.19785** (2023), https://api.semanticscholar.org/CorpusID:264805345 3
21. Lee, I., Park, W., Jang, J.W., Noh, M., Shim, K., Shim, B.: Towards comprehensive scene understanding: Integrating first and third-person views for lvlms. ArXiv **abs/2505.21955** (2025), https://api.semanticscholar.org/CorpusID:278959196 3, 4
22. Lee, I., Park, W., Jang, J., Noh, M., Shim, K., Shim, B.: Towards comprehensive scene understanding: Integrating first and third-person views for lvlms. arXiv preprint arXiv:2505.21955 (2025) 2
23. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer (2024) 9
24. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models (2024) 9
25. Li, Y., Zhang, Y., Lin, T., Liu, X., Cai, W., Liu, Z., Zhao, B.: Sti-bench: Are mllms ready for precise spatial-temporal world understanding? ArXiv **abs/2503.23765** (2025), https://api.semanticscholar.org/CorpusID:277452593 3
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 2
27. Liu, J., Liu, Z., Cen, Z., Zhou, Y., Zou, Y., Zhang, W., Jiang, H., Ruan, T.: Can multimodal large language models understand spatial relations? In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 620–632 (2025) 3, 4
28. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9264–9275 (2023), https://api.semanticscholar.org/CorpusID:257631738 4
29. Liu, Y., Zhang, B., Zang, Y., Cao, Y., Xing, L., wen Dong, X., Duan, H., Lin, D., Wang, J.: Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. ArXiv **abs/2510.27606** (2025), https://api.semanticscholar.org/CorpusID:282719317 3, 9

30. Lu, H., Liu, W., Zhang, B., Wang, B.L., Dong, K., Liu), B.L.B., Sun, J., Ren, T., Li, Z., Yang, H., Sun, Y., Deng, C., Xu, H., Xie, Z., Ruan, C.: Deepseek-vl: Towards real-world vision-language understanding. ArXiv **abs/2403.05525** (2024), https://api.semanticscholar.org/CorpusID:268297008 9
31. Ma, C., Lu, K., Cheng, T.Y., Trigoni, N., Markham, A.: Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. Advances in Neural Information Processing Systems 37 (2024), https://api.semanticscholar.org/CorpusID:268536785 4
32. Ma, Y., Song, Z., Zhuang, Y., Hao, J., King, I.: A survey on vision-language-action models for embodied ai. arXiv preprint arXiv:2405.14093 (2024) 1
33. Marsili, D., Agrawal, R., Yue, Y., Gkioxari, G.: Visual agentic ai for spatial reasoning with a dynamic api. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19446–19455 (2025), https://api.semanticscholar.org/CorpusID:276249710 4
34. Meng, Z., Zhou, H., Chen, Y.: I know about "up"! enhancing spatial reasoning in visual language models through 3d reconstruction. ArXiv **abs/2407.14133** (2024), https://api.semanticscholar.org/CorpusID:271310255 4
35. OpenAI: Gpt-5 system card (8 2025), system card, OpenAI, August 2025 9
36. Ravi, S., Sarch, G., Vineet, V., Wilson, A.D., Kumaravel, B.T.: Out of sight, not out of context? egocentric spatial reasoning in vlms across disjoint frames. ArXiv **abs/2505.24257** (2025), https://api.semanticscholar.org/CorpusID:279070962 4
37. Ray, A., Duan, J., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., Zeng, K.H., Saenko, K.: Sat: Dynamic spatial aptitude training for multimodal language models (2024), https://api.semanticscholar.org/CorpusID:274609987 2
38. Reilly, D., Govind, M.K., Xue, L., Das, S.: From my view to yours: Ego-to-exo transfer in vlms for understanding activities of daily living. arXiv preprint arXiv:2501.05711 (2025) 3, 4
39. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7396–7404 (2018) 1
40. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.T.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15768–15780 (2024), https://api.semanticscholar.org/CorpusID:274233932 2, 3
41. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11888–11898 (2023) 2
42. Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., et al.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025), https://arxiv.org/abs/2507.01006 9
43. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024) 3, 4
44. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In: European conference on computer vision. pp. 605–621. Springer (2020) 1

45. Wolbers, T., Wiener, J.M.: Challenges for identifying the neural mechanisms that support spatial navigation: the impact of spatial scale. Frontiers in human neuroscience **8**, 571 (2014) 1
46. Wu, H., Huang, X., Chen, Y., Zhang, Y., Wang, Y., Xie, W.: Spatialscore: Towards comprehensive evaluation for spatial intelligence (2025), https://api.semanticscholar.org/CorpusID:283737154 4
47. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025) 2, 3
48. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., Lin, D., Wang, T., Pang, J.: Mmsi-bench: A benchmark for multi-image spatial intelligence. ArXiv **abs/2505.23764** (2025), https://api.semanticscholar.org/CorpusID:278995731 3, 4
49. Zheng, X., Dongfang, Z., Jiang, L., Zheng, B., Guo, Y., Zhang, Z., Albanese, G., Yang, R., Ma, M., Zhang, Z., Liao, C., Zhen, D., Lyu, Y., Fu, Y., Ren, B., Zhang, L., Paudel, D.P., Sebe, N., van Gool, L., Hu, X.: Multimodal spatial reasoning in the large model era: A survey and benchmarks. ArXiv **abs/2510.25760** (2025), https://api.semanticscholar.org/CorpusID:282575141 4
50. Zhou, Q., Zhou, R., Hu, Z.Y., Lu, P., Gao, S., Zhang, Y.: Image-of-thought prompting for visual reasoning refinement in multimodal large language models. ArXiv **abs/2405.13872** (2024), https://api.semanticscholar.org/CorpusID:269982092 4
51. Zhu, F., Wang, H., Xie, Y., Gu, J., Ding, T., Yang, J., Jiang, H.: Struct2d: A perception-guided framework for spatial reasoning in mllms (2025), https://api.semanticscholar.org/CorpusID:282758225 4
52. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. ArXiv **abs/2504.10479** (2025), https://api.semanticscholar.org/CorpusID:277780955 9