Can VLMs Handle Multi-hop Compositional Spatial Reasoning?

Youngwan Lee[1,2*] Soojin Jang[1*] Yoorhim Cho[1] Seunghwan Lee[3]
Yong-Ju Lee[1] Sung Ju Hwang[2,4]

[1]ETRI [2]KAIST AI [3]Sungkyunkwan University [4]DeepAuto.ai

Abstract

Spatial reasoning is a critical capability for Vision–Language Models (VLMs), particularly when deployed as Vision–Language–Action (VLA) agents in real-world environments. However, existing benchmarks predominantly focus on simple, single-hop spatial questions, falling short of capturing the multi-hop reasoning and precise visual grounding required in practical scenarios. To address this gap, we introduce MultihopSpatial, a benchmark designed for multi-hop compositional spatial reasoning with 1–3 hop questions across ego- and exo-centric perspectives. Through extensive evaluation of 30 state-of-the-art VLMs, we demonstrate that compositional spatial reasoning remains a significant challenge for current VLMs.

1. Introduction

Recent interest in physical AI has accelerated the development of embodied agents, particularly Vision-Language-Action (VLA) models. These agents rely heavily on Vision-Language Models (VLMs) for spatial reasoning in real-world environments. However, current VLMs often lack precise visual grounding, limiting their ability to support accurate perception and action. In practice, agents must perform multi-step spatial reasoning and accurately localize target objects to complete tasks reliably. For example, an instruction such as "*Could you move the round cup on my right—the one furthest away?*" requires the agent to adopt an ego-centric perspective, identify the relevant area(*right*), filter objects by attributes(*round*), and compare spatial relations(*furthest*)—a process that closely resembles multi-hop reasoning combined with precise bounding box prediction. Ultimately, successful navigation and manipulation depend on both correct reasoning and accurate visual grounding.

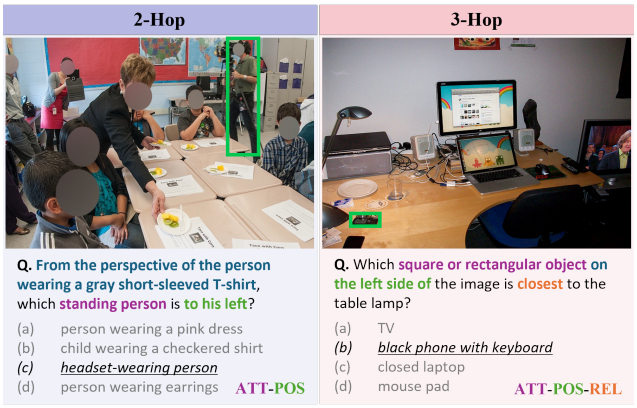

Figure 1. **Example MultihopSpatial questions for 2-hop (ego) and 3-hop (exo) reasoning.** We omit the phrase "`and provide the bounding box coordinate of the region related to your answer`" for brevity.

VLM spatial reasoning benchmarks have rapidly evolved from elementary relations [7, 9] to diverse dimensions, including 3D properties [16], video [24], scale [23], real-world complexity [22], fine-grained taxonomies [12], multi-image contexts [26], and perspectives [15]. However, these benchmarks predominantly rely on single-hop queries without requiring explicit target localization. As a result, they under-evaluate the compositional reasoning and visual grounding essential for real-world embodied scenarios.

To bridge this gap, we introduce **MultihopSpatial**, a comprehensive benchmark for evaluating multi-hop, compositional spatial reasoning paired with visual grounding. It comprises 4,500 QA pairs spanning 1- to 3-hop complexities across attribute, position, and relation conditions, encompassing both ego- and exo-centric perspectives to mirror real-world interactions. Crucially, MultihopSpatial advances beyond standard multiple-choice evaluation by requiring models to localize the target object via bounding box prediction, thereby assessing whether a model truly

*Equal contribution

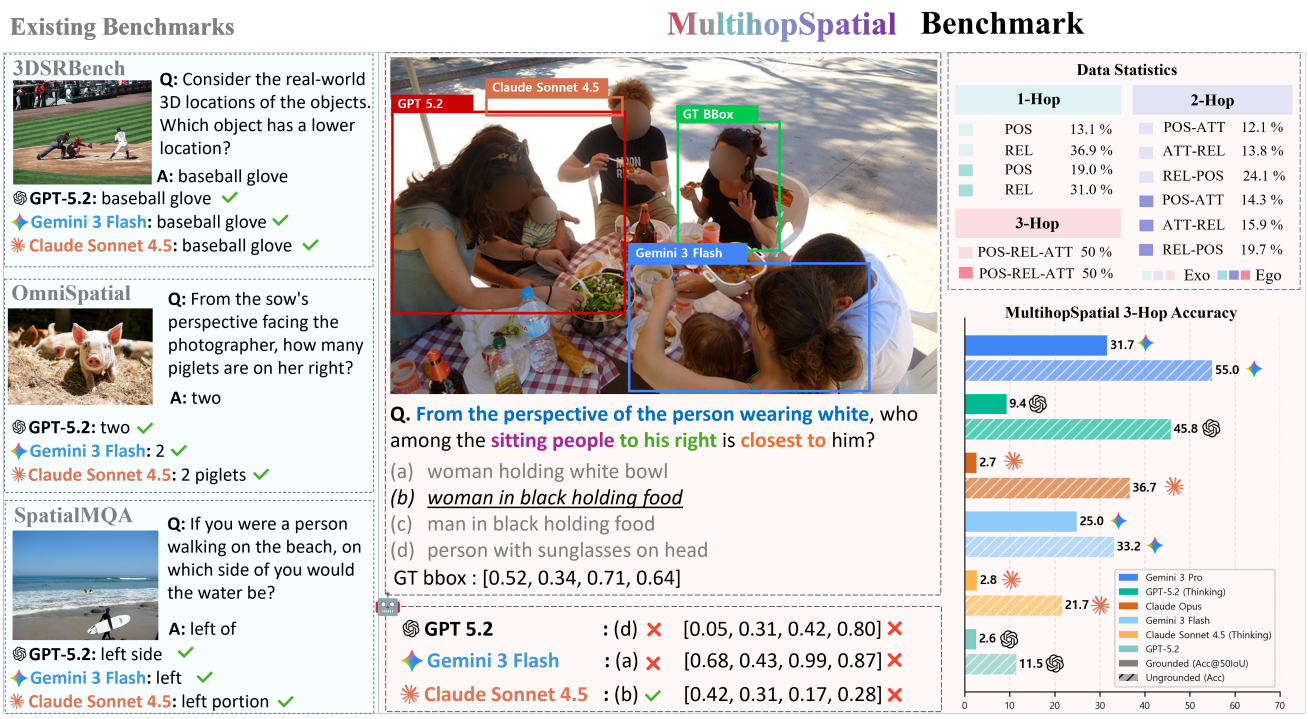

Figure 2. **Comparison of existing benchmarks and MultihopSpatial benchmark.** In the question text, colored spans denote the queried reasoning components: Perspective, Attribute, Position, and Relation.

grounds its reasoning in the visual scene rather than arriving at correct answers through superficial shortcuts. Through extensive evaluation of 30 VLMs—including state-of-the-art commercial, open-weight, and specialized spatial reasoning models—we reveal that compositional spatial reasoning remains a formidable challenge. For instance, even a highly capable reasoning model such as GPT-5.2-Thinking achieves only 45.8% accuracy on 3-hop questions, with performance degrading substantially when grounding is taken into account (Fig. 2). These findings underscore a major gap between current VLM capabilities and the compositional spatial understanding required for real-world applications.

2. MultihopSpatial

Data Source, Annotation and Verification. We curate 3,563 spatially complex images from COCO [14] and PACO-Ego4D [20], ensuring diverse coverage of everyday indoor/outdoor scenes and ego/exo perspectives. Upon these images, we construct 4,500 multiple-choice questions (MCQs), perfectly balanced across 1- to 3-hop reasoning levels (1,500 per hop) and viewpoints (750 ego-centric and 750 exo-centric per hop).

To completely eliminate the reliability concerns and hallucinations inherent in AI-generated data, all QA pairs and bounding boxes were annotated by ten trained human experts. Each sample underwent a rigorous multi-stage verification process with three rounds of independent cross-checking. Finally, three verifiers ensured that: (i) all option entities exist in the image, (ii) the bounding-box annotation precisely matches the referred target, and (iii) the labeled answer is correct and uniquely supported by the question. This protocol yields a high inter-annotator agreement (Krippendorff's $\alpha = 0.90$), ensuring a high-quality and dependable benchmark for evaluating complex spatial reasoning.

Spatial Reasoning Categories. We define three spatial reasoning categories: ATTRIBUTE (ATT), covering visual properties; POSITION (POS), referring to spatial location and orientation; and RELATION (REL), capturing spatial relationships. Our benchmark composes them into multi-hop questions (*e.g.*, 2- and 3-hop) that require sequential inferences. Higher hop counts introduce longer reasoning chains and greater difficulty, enabling fine-grained diagnosis across reasoning complexity levels. The distribution across hop counts and categories is summarized in the upper-right panel of Fig. 2.

1-Hop. Single-step questions target a single spatial category (POS or REL). We exclude ATT as a standalone category, since attributes are mainly perceptual unless combined with spatial metrics. Although prior work has explored 1-hop spatial reasoning [12, 15, 16, 22], we include it as a controlled baseline for depth-wise comparison with

Table 1. **Benchmark results across different hop counts and Ego/Exo perspectives**. Blue cells indicate the best performance within each model group.

Model	Overall			3Hop-Ego		3Hop-Exo		2Hop-Ego		2Hop-Exo		1Hop-Ego		1Hop-Exo	
	Acc.	Acc@50	avg. IoU	Acc.	Acc@50	Acc.	Acc@50	Acc.	Acc@50	Acc.	Acc@50	Acc.	Acc@50	Acc.	Acc@50
Proprietary Models – Instant															
Claude-Opus-4.5 [3]	45.1	3.2	13.3	25.7	2.0	48.5	3.6	33.7	2.0	58.1	4.8	43.2	3.5	61.1	3.1
Claude-Sonnet-4.5 [4]	20.9	0.5	4.2	6.4	0.4	18.5	0.9	12.3	0.0	34.7	1.6	13.5	0.0	40.0	0.1
GPT-5.2 [19]	19.3	2.0	11.8	5.3	0.7	18.0	4.9	10.8	0.1	30.4	5.9	8.1	0.0	43.3	0.4
Open-weight Models – Instant															
Qwen3-VL-235B-Instruct [5]	41.3	34.8	71.1	14.8	12.3	42.9	37.6	21.9	18.5	58.5	52.7	30.7	23.5	79.2	64.4
Qwen3-VL-32B-Instruct [5]	40.9	33.4	69.6	13.9	9.7	43.9	36.4	21.9	17.2	59.2	53.1	30.4	22.3	76.4	61.6
Qwen3-VL-8B-Instruct [5]	38.0	31.3	69.5	12.3	8.8	42.1	36.1	18.8	15.2	55.3	49.1	26.7	20.8	72.8	58.0
InternVL-3.5-38B [21]	40.8	9.7	28.7	17.5	3.2	44.5	12.3	24.3	4.5	56.8	24.5	31.9	3.3	69.6	10.4
InternVL-3.5-14B [21]	39.7	7.9	26.2	17.1	2.5	44.5	10.1	22.0	2.8	56.1	14.4	30.4	3.3	68.0	14.0
Gemma-3-IT-27B [13]	33.1	0.4	5.4	18.1	0.1	30.8	0.5	22.0	0.1	45.7	1.1	28.1	0.3	53.9	0.3
Gemma-3-IT-12B [13]	29.8	0.4	5.9	16.9	0.3	31.6	0.5	22.1	0.4	40.8	0.5	22.9	0.1	44.5	0.5
GLM-4.6V [10]	43.2	35.2	69.5	15.9	12.3	46.7	39.3	22.7	18.4	61.6	53.2	32.4	24.3	80.1	63.7
Molmo2-8B [8]	41.8	0.3	8.8	15.9	0.3	44.4	0.4	21.6	0.3	60.4	0.1	32.4	0.4	76.4	0.3
Proprietary Models – Reasoning															
Gemini-3-Pro [11]	64.7	40.6	55.0	39.7	18.8	71.1	45.3	36.8	20.5	81.2	55.5	71.1	41.1	88.4	62.3
GPT-5.2-Thinking [19]	57.9	11.5	29.0	36.1	8.5	55.7	10.4	49.7	7.6	63.6	18.0	65.6	12.5	76.4	11.7
Gemini-3-Flash [11]	57.2	40.2	61.2	6.9	4.3	61.2	46.9	42.3	25.3	80.0	63.7	66.0	38.9	86.8	62.1
Claude-Opus-4.5-Thinking [3]	47.0	4.7	16.7	25.5	3.5	49.7	4.7	35.2	3.1	60.0	8.8	45.1	5.2	66.5	2.9
Claude-Sonnet-4.5-Thinking [4]	32.2	4.3	19.2	14.7	1.9	29.9	3.6	22.1	2.3	45.7	8.1	31.3	3.9	49.3	6.1
Open-weight Models – Reasoning															
Qwen3-VL-235B-Thinking [5]	45.1	36.3	67.8	17.6	12.7	51.2	42.3	24.8	19.3	67.6	58.7	31.2	22.8	78.1	61.9
Qwen3-VL-32B-Thinking [5]	46.8	37.4	67.2	19.2	12.9	57.5	47.1	24.3	18.1	70.1	60.0	30.4	23.1	79.6	63.2
Qwen3-VL-8B-Thinking [5]	41.7	29.5	60.1	18.5	9.2	47.9	36.4	21.3	11.9	63.3	51.6	28.5	16.5	70.5	51.2
InternVL-3.5-38B-Thinking [21]	42.1	27.4	57.0	19.5	10.9	43.3	32.5	24.7	15.3	56.8	39.2	34.8	20.7	73.6	45.9
InternVL-3.5-14B-Thinking [21]	38.2	11.1	34.7	14.1	4.7	42.8	13.6	21.1	5.6	55.3	20.5	27.6	6.4	68.0	15.9
GLM-4.6V-Thinking [10]	42.0	34.7	70.1	14.1	10.7	46.3	40.0	19.5	15.3	63.1	56.1	31.2	23.3	77.6	62.7
Specialized Spatial Reasoning Models															
SenseNova-InternVL3-8B [6]	42.3	17.3	38.8	20.4	9.1	45.2	19.7	25.2	9.2	55.5	27.2	34.5	11.2	73.2	27.2
Cosmos-Reason2-8B [18]	37.8	27.9	61.4	15.2	10.5	40.7	31.9	19.5	13.5	54.5	43.5	26.5	17.1	70.1	51.1
VST-7B-RL [25]	36.0	0.0	1.5	16.7	0.0	34.1	0.1	23.9	0.0	48.1	0.0	24.5	0.0	68.8	0.0
SpaceQwen3-VL-2B [7]	33.6	10.1	31.5	18.5	4.0	32.5	9.9	22.8	4.0	47.2	22.9	26.1	4.4	54.3	15.2
SpaceOm [1]	32.3	0.3	2.6	15.3	0.5	37.9	0.1	19.6	0.1	47.9	0.4	20.5	0.3	52.8	0.4
SpatialReasoner [17]	31.7	8.7	29.9	18.0	6.8	34.0	10.3	19.6	4.3	46.0	13.7	21.3	5.7	51.5	11.2
SpaceThinker-3B [2]	31.1	4.0	16.6	15.9	2.9	36.3	3.7	19.2	2.9	44.5	6.5	20.8	3.3	50.0	4.3

multi-hop compositions.

2-Hop. Questions combining two categories (ATT+POS, ATT+REL, or POS+REL). As shown in Fig. 1, they typically follow a two-stage structure: (i) restricting the candidate set using one category and (ii) identifying the target using the other. We define this as "2-hop" because both constraints must be satisfied, regardless of inference order.

3-Hop. Questions incorporating all three categories (ATT+POS+REL) in a single query. An ATT cue narrows candidates, after which the model reasons over POS and REL to identify the target (*e.g.*, selecting the rightmost object that is farthest/closest). This structure (Fig. 1) mirrors how humans refer to objects in cluttered scenes, testing the disambiguation needed for embodied task execution.

3. Experiments

3.1. Experiment Setup

We benchmark 30 VLMs on MultihopSpatial, spanning five categories: (i) **Proprietary instant models**: Claude-Opus-4.5, Claude-Sonnet-4.5, and GPT-5.2; (ii) **Proprietary reasoning models**: Gemini-3-Pro[1] &-Flash, GPT-5.2-Thinking (xhigh), Claude-Opus-4.5-Thinking, and Claude-Sonnet-4.5-Thinking; (iii) **Open-weight instant models**: Qwen3-VL-Instruct (8B,32B,235B-A22B), InternVL-3.5 (14B,38B), GLM-4.6V, Gemma-3-IT (12B,27B), and Molmo2-8B; (iv) **Open-weight reasoning models**: thinking-mode variants of Qwen3-VL (32B,235B-A22B), InternVL-3.5 (14B,38B), and GLM-4.6V. (v) **Spatial reasoning model**: SenseNova-SI-1.3-InternVL3-8B [6], Cosmos-Reason2-8B [18], VST-7B-RL [25], SpaceQwen3-VL-2B [7], SpaceOm [1], SpatialReasoner [17], and SpaceThinker2.5VL-3B [2]. All models are prompted with the same template and required to provide both a multiple-choice answer and a bounding box prediction.

3.2. Evaluation Metric

MCQ Accuracy. Measures the percentage of correct multiple-choice predictions ($\hat{y} = y^*$). While standard, it does not verify spatial localization.

Acc@50IoU. Our primary grounded metric requires correct

[1] Gemini-3-Pro&-Flash operate with thinking mode enabled by default. We therefore classify them as reasoning models.

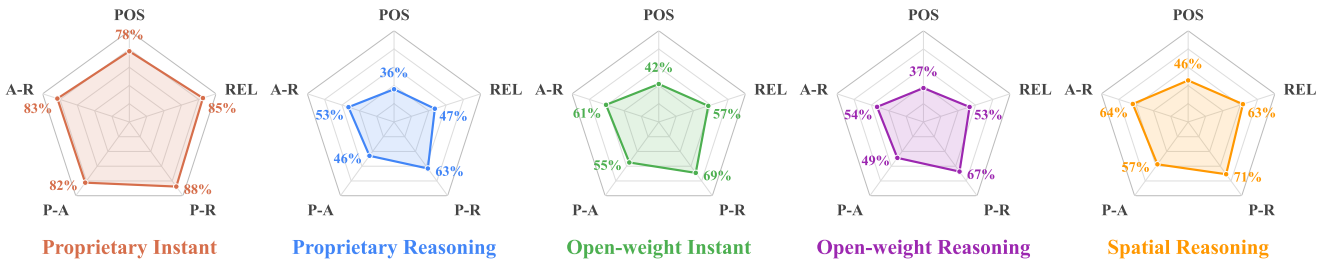

Figure 3. **Average Error Rates (↓, %) by tag combination across model categories.** A, R, and P denote Attribute, Relation, and Position.

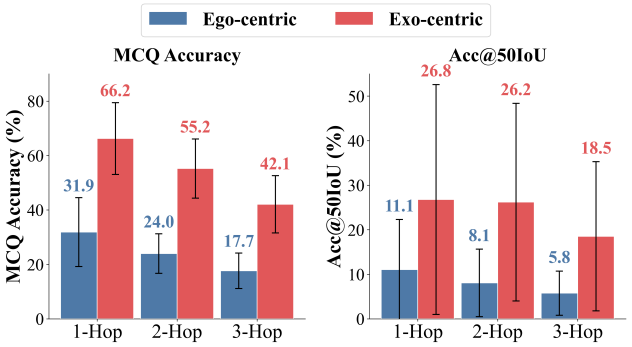

Figure 4. **Average Performance by Hop Count.** MCQ accuracy and Acc@50IoU across 1- to 3-hop queries for ego- and exo-centric perspectives.

answer selection and precise localization. A prediction is correct only if $\hat{y} = y^*$ and $\text{IoU}(\hat{B}, B^*) \geq 0.5$. This filters out ungrounded predictions, ensuring genuine localization.

3.3. Main Results

Overall Performance Highlights. As shown in Tab. 1, Gemini-3-Pro [11] achieves the highest MCQ accuracy and Acc@50IoU, while Qwen3-VL-32B-Thinking [5] leads among open-weight models. However, strong answer selection does not necessarily correspond to accurate localization, as evidenced by Qwen3-VL-235B [5] achieving the highest Avg IoU. These results suggest that reasoning and grounding remain imperfectly aligned in current VLMs.

Metric-dependent Rankings. This decoupling causes rank reversals between metrics. Models like Claude-Opus-4.5 [4] and Molmo2-8B [8] rank high in MCQ but plummet in Acc@50IoU (*e.g.*, Claude [4] drops from 7th to 29th), indicating shortcut-based answers without genuine localization. These shifts show MCQ alone is misleading, making Acc@50IoU essential to verify true spatial understanding.

Benchmark Difficulty. With the best model peaking at just 40.6% Acc@50IoU, MultihopSpatial remains far from saturated. The difficulty peaks under 3-hop ego-centric conditions, where only 3 of 37 models exceed the 25% random MCQ baseline, and just 9 surpass 10% Acc@50IoU. Strikingly, even advanced reasoning models like GPT-5.2-Thinking [19] (8.5%) and Claude-Sonnet-4.5-Thinking [4](1.9%) fail drastically here, confirming our benchmark rigorously evaluates both compositional reasoning and spatial grounding capabilities of current VLMs.

3.4. Additional Analysis

Ego vs. Exo: Perspective-Taking as a Compounding Bottleneck. As shown in Fig. 4, exo-centric queries consistently outperform ego-centric ones across all hop counts and metrics. In particular, ego-centric performance degrades more sharply with increasing hops, especially under Acc@50IoU, where it drops to 5.8% at 3-Hop. These results suggest that ego-centric perspective-taking introduces compounding errors across reasoning hops, making it a key bottleneck for compositional spatial reasoning.

Error Analysis on Tag Compositions. As shown in Fig. 3, reasoning models consistently outperform instant models, yet multi-tag compositions remain a major bottleneck. In particular, the POS-REL (P-R) setting yields much higher error rates than single-tag cases, indicating the difficulty of handling positional localization and relational comparison. Even specialized spatial reasoning models still struggle on these compositions, suggesting that compositional spatial reasoning remains unresolved for current VLMs.

4. Conclusion

We introduce **MultihopSpatial**, a benchmark for evaluating multi-hop compositional spatial reasoning with visual grounding in VLMs. Our benchmark requires models to jointly perform compositional reasoning and precise target localization via bounding box prediction. Through extensive evaluation of 30 VLMs, we reveal that compositional spatial reasoning remains a significant challenge for current models. We hope our dataset will catalyze future research on advancing spatial intelligence in VLMs.

Acknowledgments This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration).

References

[1] Remyx AI. Spaceom: A multimodal model for space exploration. https://huggingface.co/remyxai/SpaceOm, 2024. Accessed: 2026-03-02. 3

[2] Remyx AI. Spacethinker-qwen2.5vl-3b. https://huggingface.co/remyxai/SpaceThinker-Qwen2.5VL-3B, 2025. Accessed: 2026-03-02. 3

[3] Anthropic. Claude opus 4.5 system card. Technical report, Anthropic, 2025. Accessed: 2026-03-02. 3

[4] Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic, 2025. Accessed: 2026-03-02. 3, 4

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 3, 4

[6] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025. 3

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1, 3

[8] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026. 3, 4

[9] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, pages 148–166. Springer, 2024. 1

[10] GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 3

[11] Google. Gemini 3. https://gemini.google.com/, 2025. Accessed: 2025-03-02. 3, 4

[12] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *ICLR*, 2026. 1, 2

[13] Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4, 2025. 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2

[15] Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In *ACL*, pages 620–632, Vienna, Austria, 2025. Association for Computational Linguistics. 1, 2

[16] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *ICCV*, pages 6924–6934, 2025. 1, 2

[17] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. 2025. 3

[18] NVIDIA. Cosmos-reason2-8b. https://huggingface.co/nvidia/Cosmos-Reason2-8B, 2024. Accessed: 2025-03-02. 3

[19] OpenAI. Update to GPT-5 system card: GPT-5.2. Technical report, OpenAI, 2025. Accessed: 2026-03-02. 3, 4

[20] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, pages 7141–7151, 2023. 2

[21] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3

[22] Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md Iqramul Hoque, et al. Spatialab: Can vision-language models perform spatial reasoning in the wild? *arXiv preprint arXiv:2602.03916*, 2026. 1, 2

[23] Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards unified evaluation for multimodal spatial understanding. *arXiv e-prints*, pages arXiv–2505, 2025. 1

[24] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, pages 10632–10643, 2025. 1

[25] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025. 3

[26] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 1