MultihopSpatial: Multi-hop Compositional Spatial Reasoning Benchmark for Vision-Language Model

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

[1] Electronics and Telecommunications Research Institute, South Korea
[2] Korea Advanced Institute of Science and Technology, South Korea
[3] Sungkyunkwan University, South Korea
[4] DeepAuto, South Korea

Website Dataset Model Code

Abstract. Spatial reasoning is foundational for Vision-Language Models (VLMs), particularly when deployed as Vision-Language-Action (VLA) agents in physical environments. However, existing benchmarks predominantly focus on elementary, single-hop relations, neglecting the multi-hop compositional reasoning and precise visual grounding essential for real-world scenarios. To address this, we introduce **MultihopSpatial**, offering three key contributions: (1) A comprehensive benchmark designed for multi-hop and compositional spatial reasoning, featuring 1- to 3-hop complex queries across diverse spatial perspectives. (2) **Acc@50IoU**, a complementary metric that simultaneously evaluates reasoning and visual grounding by requiring both answer selection and precise bounding box prediction—capabilities vital for robust VLA deployment. (3) **MultihopSpatial-Train**, a dedicated large-scale training corpus to foster spatial intelligence. Extensive evaluation of 37 state-of-the-art VLMs yields eight key insights, revealing that compositional spatial reasoning remains a formidable challenge. Finally, we demonstrate that reinforcement learning post-training on our corpus enhances both intrinsic VLM spatial reasoning and downstream embodied manipulation performance.

1 Introduction

The recent surge of interest in physical AI has accelerated the development of embodied agents, particularly Vision-Language-Action (VLA) models [7,21,35,39]. These agents fundamentally rely on Vision-Language Models (VLMs) for spatial reasoning to interact with the real world. However, existing VLMs often lack precise visual grounding, leading to significant difficulties in accurate perception and action execution within VLA frameworks [10,18,43,44]. In complex environments, agents must execute multi-step, compositional spatial reasoning and

* Equal contribution
** Corresponding author

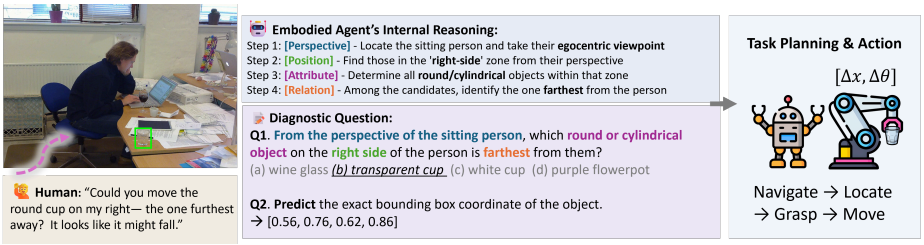


Fig. 1: A multi-hop spatial reasoning example for an embodied agent in the real-world scenario.

accurately localize targets to succeed. As illustrated in Fig. 1, an instruction like "*Could you move the round cup on my right—the one furthest away?*" requires an agent to adopt an ego-centric perspective, isolate the target position, filter by attributes, and compare relations. This internal reasoning perfectly mirrors a diagnostic multi-hop multiple-choice question (MCQ) coupled with exact bounding box prediction. Ultimately, an agent can only successfully navigate and manipulate an object if it accurately answers the query and visually grounds the target.

Despite rapid advancements, existing spatial reasoning benchmarks largely focus on single-hop queries, such as verifying elementary relations like "Is X to the right of Y?" [13,19,24,26,36]. As depicted in the left panel of Fig. 2, these datasets typically evaluate spatial understanding through standard MCQs without requiring spatial localization. This design not only fails to capture the compositional reasoning required in dynamic physical environments but also leaves a significant spatial blind spot in VLM evaluation [3], as models can often select the correct answer without genuinely locating the target. Furthermore, while some recent efforts provide training sets alongside their benchmarks [19,38], they typically do not extend their validation to end-to-end VLA execution, leaving a profound gap between VLM spatial reasoning metrics and actual embodied action performance.

To bridge this gap, we introduce **MultihopSpatial**, a comprehensive benchmark 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 dimensions, encompassing both ego-centric and exo-centric viewpoints to mirror real-world interactions. To ensure robust evaluation for VLA deployment, we introduce a complementary grounded metric, **Acc@50IoU**, which considers an answer correct only if the predicted bounding box overlaps the ground-truth by at least 50% IoU. The critical necessity of this metric is starkly illustrated in Fig. 2 (right): while frontier VLMs achieve superficially high standard MCQ accuracy (solid bars), their performance plummets under Acc@50IoU (striped bars). This severe discrepancy confirms that standard MCQs mask a profound lack of spatial grounding, validating Acc@50IoU as an essential tool to expose genuine capabilities. Finally, we provide a 6,791-sample training set, validating the dataset's utility not merely as an evaluation tool, but as a potent corpus for enhancing downstream VLA performance.

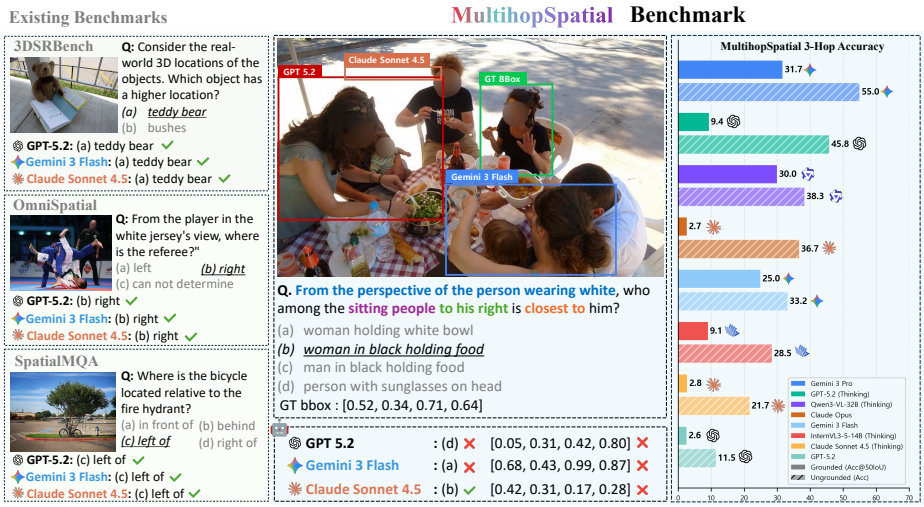


Fig. 2: Comparison of existing benchmarks (single-hop) and MultihopSpatial benchmark. In the question text, colored spans denote the queried reasoning components: Perspective, Attribute, Position, and Relation.

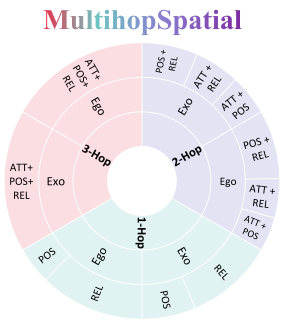
We extensively evaluate 37 VLMs, encompassing state-of-the-art commercial, open-weight, and specialized spatial reasoning models. Our findings reveal that multi-hop spatial reasoning remains a formidable challenge; for instance, a highly capable reasoning model (*e.g.*, GPT-5.2-Thinking) drops from 45.8% in standard MCQ to a mere 9.4% under our grounded metric on MultihopSpatial 3-Hop in Fig. 2 (right). Beyond evaluation, we leverage the MultihopSpatial-Train set to post-train a base VLM using reinforcement learning (*e.g.*, GRPO [33]). Our experiments demonstrate that this RL post-training not only enhances the model's intrinsic multi-hop spatial reasoning capabilities across five benchmarks but also translates to improved performance in two downstream embodied VLA tasks.

Our main contributions are summarized as follows:

- **MultihopSpatial:** A novel benchmark jointly evaluating multi-hop, compositional spatial reasoning and visual grounding in VLMs.
- **Acc@50IoU:** A complementary grounded metric requiring both correct answer selection and precise bounding box prediction, eliminating the blind spot of traditional MCQs.
- **MultihopSpatial-Train:** A dedicated training corpus where RL post-training enhances both intrinsic VLM spatial intelligence and downstream VLA manipulation performance.

2 Related Work

VLM spatial reasoning benchmarks have rapidly evolved from elementary relations (SpatialVLM [9], BLINK [14]) to diverse dimensions, including 3D prop-



1-Hop		2-Hop		3-Hop	
POS	13.1 %	POS-ATT	12.1 %	POS-REL-ATT	50 %
REL	36.9 %	ATT-REL	13.8 %	POS-REL-ATT	50 %
POS	19.0 %	REL-POS	24.1 %		
REL	31.0 %	POS-ATT	14.3 %		
		ATT-REL	15.9 %		
		REL-POS	19.7 %		

Exo-centric View
Ego-centric View

Category	Description
Attribute (ATT)	Visual properties (Texture, Shape, Color, Pattern, Material, etc.)
Position (POS)	Spatial location and orientation (Left, Right, Behind, In front of, etc.)
Relation (REL)	Spatial relationship (Close, Far, High, Low, etc.)

Fig. 3: Compositional structure and category definitions in MultihopSpatial.

erties (3DSRBench [26]), video (VSI-Bench [40]), scale (SpatialScore [38]), real-world complexity (SpatiaLab [36]), fine-grained taxonomies (OmniSpatial [19]), multi-image contexts (MMSI-Bench [42]), and perspectives (SpatialMQA [24]). However, they predominantly rely on single-hop queries, under-evaluating the compositional reasoning essential for real-world scenarios. MultihopSpatial addresses this by introducing 1- to 3-hop complex queries. Crucially, we advance beyond standard MCQs with a grounded metric requiring both correct answer selection and precise spatial localization, ensuring evaluations explicitly reflect the visual grounding vital for VLA deployment.

3 MultihopSpatial

3.1 Overview

We introduce **MultihopSpatial**, a 4.5K-sample benchmark of manually annotated VQA pairs focusing on multi-hop, compositional spatial queries common in real-world scenarios. Each example provides ground-truth bounding boxes to jointly evaluate reasoning and spatial localization, which is crucial for embodied/VLA settings (*e.g.*, grasping and action). This grounded design improves interpretability and eliminates the evaluation blind spot of random guessing.

3.2 Dataset Construction

Data Source, Annotation and Verification. We curate 3,563 spatially complex images from COCO [22] and PACO-Ego4D [31], ensuring diverse coverage of everyday indoor/outdoor scenes and ego/exo perspectives. Upon these images, we construct **4,500** multiple-choice questions (MCQs) for the proposed **MultihopSpatial benchmark**, 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 strictly annotated by ten trained human experts. Each sample underwent a rigorous multi-stage

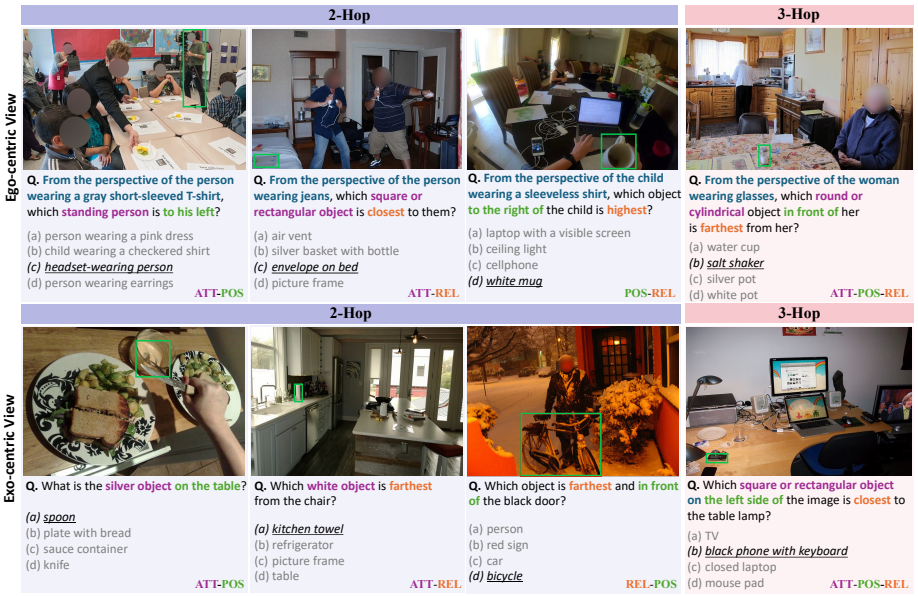


Fig. 4: Example MultihopSpatial questions for 2-hop and 3-hop reasoning under ego-centric (top) and exo-centric (bottom) views. We omit the phrase "`and provide the bounding box coordinate of the region related to your answer`" for brevity.

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 strict protocol yielded a high inter-annotator agreement (Krippendorff's $\alpha = 0.90$), ensuring a high-quality and dependable benchmark for evaluating complex spatial reasoning. Full annotation guidelines and verification procedures are detailed in Appendix E. Furthermore, we introduce **MultihopSpatial-Train**, an auxiliary large-scale training corpus of **6,791** grounded VQA samples designed to support the post-training of VLMs for spatial intelligence.

Spatial Reasoning Categories. We define three fundamental spatial reasoning categories: Attribute (att), Position (pos), and Relation (rel), detailed in Fig. 3. Our benchmark systematically composes these categories into multi-hop questions (*e.g.*, 2- and 3-hop) that demand sequential intermediate inferences, *e.g.*, identifying candidates, restricting the search space, comparing relations, and resolving ambiguities. This compositional structure makes each intermediate output constrain the next step's search space, rather than treating all conditions as a single parallel conjunctive filter. Such compositional reasoning is essential for robust real-world scene understanding and embodied interaction. By increasing the hop count, we represent longer reasoning chains and escalating difficulty, enabling fine-grained diagnosis of model performance across different levels of reasoning complexity.

1-Hop. Single-step questions targeting one spatial category (POS or REL). We exclude ATT as a standalone category, since attributes are primarily perceptual unless composed with spatial metrics. While 1-hop spatial reasoning has been explored in prior works [19, 24, 26, 36], we deliberately include it to serve as a controlled baseline for depth-wise comparisons against multi-hop compositions.

2-Hop. Questions combining two categories (ATT+POS, ATT+REL, or POS+REL). As shown in Fig. 4, these queries typically follow a two-stage structure: (i) restricting the candidate set using one category, and (ii) identifying the target by applying the other. We define this as 2-hop" because both constraints must be jointly satisfied, regardless of the specific inference order.

3-Hop. Questions incorporating all three categories (ATT+POS+REL) in a single query. An ATT cue narrows candidates, POS further restricts the spatially valid subset, and REL selects the final target within that subset (*e.g.*, selecting the rightmost object that is farthest/closest). This structure (Fig. 4) mirrors how humans refer to objects in cluttered scenes, testing the disambiguation needed for embodied task execution.

4 Reinforcement Learning on MultihopSpatial-Train

Beyond serving as a static evaluation benchmark, we investigate the utility of MultihopSpatial as a training corpus. To validate whether it can effectively enhance the multi-hop spatial reasoning capabilities of VLMs, we post-train a base VLM (*e.g.*, Qwen3-VL-4B-Instruct [6]) via reinforcement learning. We adopt the Reinforcement Learning with Verifiable Rewards (RLVR) paradigm [12, 33], employing deterministic, rule-based reward functions to provide unambiguous correctness signals. Our dataset is naturally suited for RLVR, as each sample provides a unique multiple-choice answer and a ground-truth bounding box for binary verification and continuous localization evaluation, respectively.

To optimize the VLM, we employ Group Relative Policy Optimization (GRPO), an efficient RL algorithm that eliminates the need for a critic model. For a given input, GRPO samples a group of responses and optimizes the policy using advantage scores obtained by normalizing sequence-level rewards within the group. This optimization is driven by our composite reward function, which jointly evaluates format adherence, answer correctness, and spatial localization accuracy.

The model is prompted to output a specific format: "`Answer: (X)`" followed by "`Bounding Box: [x1, y1, x2, y2]`", where `X` $\in \{a, b, c, d\}$ and coordinates are normalized to a $[0, 1000]$ scale [6, 16]. The total reward R for a generated sequence is simply defined as the sum of three components:

$$R = R_{\text{format}} + \alpha \cdot R_{\text{mcq}} + \beta \cdot R_{\text{bbox}}, \tag{1}$$

where we set weighting coefficients $\alpha = \beta = 1$.

Format Reward (R_{format}). A binary reward assigning 1 if the output strictly follows the expected parsing format, and 0 otherwise.

MCQ Reward (R_{mcq}). A discrete correctness signal where $R_{\text{mcq}} = 1$ if the parsed choice matches the ground-truth label, and 0 otherwise. If parsing fails, it defaults to 0.

Bounding Box Reward (R_{bbox}). To evaluate spatial grounding, we calculate the Generalized Intersection over Union (GIoU) [32] between the predicted and ground-truth bounding boxes. To provide a dense, positively-scaled training signal, we normalize the GIoU score ($\in [-1, 1]$) into a $[0, 1]$ range:

$$R_{\text{bbox}} = \frac{\text{GIoU}(\hat{B}, B^*) + 1}{2}. \tag{2}$$

A perfect overlap yields $R_{\text{bbox}} = 1$, whereas completely disjoint boxes receive $R_{\text{bbox}} < 0.5$. Unparsable bounding boxes default to 0.

By aggregating a discrete reasoning signal (R_{mcq}) with a continuous grounding signal (R_{bbox}), our composite reward design encourages the VLM to simultaneously enhance its compositional spatial reasoning and precise visual localization capabilities. Rather than optimizing solely for text-based answer correctness, this joint formulation ensures that the model's reasoning is intrinsically tied to actual visual grounding. Further training details are described in Appendix A. We emphasize that our RL-based post-training serves as a deliberately **minimal baseline** to isolate the impact of our training data, excluding the advanced techniques used by specialized spatial reasoning models [2, 27]; the reported gains thus represent a lower bound rather than a fully optimized result.

5 Experiments

5.1 Experiment Setup

We benchmark 37 VLMs spanning five categories on MultihopSpatial:
(i) **Proprietary instant models (3)**: Claude-Opus-4.5 [4],Claude–Sonnet-4.5 [5], and GPT-5.2 [30].
(ii) **Proprietary reasoning models (5)**: Gemini-3-Pro[5] &-Flash, GPT-5.2-Thinking (xhigh), Claude-Opus-4.5-Thinking, and Claude-Sonnet-4.5-Thinking.
(iii) **Open-weight instant models (13)**: Qwen3-VL-Instruct [6] (4B,8B,32B, and 235B-A22B), InternVL-3.5 [34] (4B,8B,14B,38B), GLM-4.6V [15], Gemma-3-IT [20] (4B,12B and 27B), and Molmo2-8B [11].
(iv) **Open-weight reasoning models (9)**: Qwen3-VL-Thinking (4B,8B,32B, and 235B-A22B), InternVL-3.5 with thinking mode (4B,8B,14B,38B), and GLM-4.6V, each with thinking mode enabled.
(v) **Spatial reasoning models (7)**: SenseNova-SI-1.3-InternVL3-8B [8], Cosmos-Reason2-8B [29], VST-7B-RL [41], SpaceQwen3-VL-2B [9], SpaceOm [1], SpatialReasoner [27], and SpaceThinker2.5VL-3B [2]. Reasoning variants are obtained by enabling the thinking mode of the corresponding instant model, using the `/think` tag or equivalent API option. All models are prompted with the same

[5] Gemini-3-Pro&-Flash operate with thinking mode enabled by default, and do not provide an option to fully disable it. We therefore classify them as reasoning models.

Table 1: Benchmark results across different hop counts and Ego/Exo perspectives. Green 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 [4]	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 [5]	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 [30]	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															
GLM-4.6V [15]	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 [11]	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
Qwen3-VL-235B-Instruct [6]	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 [6]	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
InternVL-3.5-38B [34]	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 [34]	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
InternVL-3.5-4B [34]	39.7	10.3	30.0	15.9	3.1	45.2	13.5	23.9	4.5	54.9	18.7	31.9	3.6	66.3	18.4
InternVL-3.5-8B [34]	38.3	9.4	30.0	14.7	2.9	41.1	10.7	23.3	4.0	50.1	17.2	30.5	4.5	70.0	17.3
Qwen3-VL-8B-Instruct [6]	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
Qwen3-VL-4B-Instruct [6]	37.8	31.0	69.9	15.5	10.9	40.1	33.7	20.9	16.0	53.5	46.9	26.7	21.2	70.3	57.2
Gemma-3-IT-27B [20]	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 [20]	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
Gemma-3-IT-4B [20]	28.4	0.2	3.0	22.1	0.0	27.2	0.3	22.5	0.1	36.8	0.3	27.1	0.3	34.7	0.1
Proprietary Models – Reasoning															
Gemini-3-Pro [16]	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 [30]	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 [16]	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 [4]	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 [5]	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-32B-Thinking [6]	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-235B-Thinking [6]	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-4B-Thinking [6]	42.6	28.7	58.3	20.4	9.2	48.1	34.8	22.3	12.0	63.9	50.7	30.8	16.3	70.4	49.2
InternVL-3.5-38B-Thinking [34]	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
GLM-4.6V-Thinking [15]	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
Qwen3-VL-8B-Thinking [6]	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-3.5-8B-Thinking [34]	40.6	5.2	22.1	20.5	1.3	39.9	6.7	24.1	1.9	57.3	10.4	30.9	3.5	70.5	7.6
InternVL-3.5-4B-Thinking [34]	40.6	4.7	21.4	18.3	1.2	41.2	4.8	24.4	1.7	58.5	9.6	32.8	1.5	68.1	9.3
InternVL-3.5-14B-Thinking [34]	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
Specialized Spatial Reasoning Models															
SenseNova-InternVL3-8B [8]	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 [29]	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 [41]	36.0	16.2	42.4	16.7	8.5	34.1	16.1	23.9	8.9	48.1	28.3	24.5	9.6	68.8	26.0
SpaceQwen3-VL-2B [9]	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	22.7	61.8	15.3	9.5	37.9	26.4	19.6	12.0	47.9	37.2	20.5	13.2	52.8	37.9
SpatialReasoner [27]	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	14.4	45.2	15.9	6.9	36.3	18.8	19.2	5.2	44.5	27.3	20.8	6.5	50.0	21.5

instruction template and evaluated under identical conditions. For each question, models are required to provide both a multiple-choice answer and a bounding box prediction for the target object (See the system prompt in Appendix D.)

Training setup. We train Qwen3-VL-4B-Instruct [6] as the base policy model using GRPO [33] with LoRA [17] applied to the LLM backbone (10 epochs, learning rate 5×10^{-5}, batch size 128). For VLA evaluation, we integrate our model into VLM4VLA [43] and train on the CALVIN ABC→D [28] and Libero [23], following VLM4VLA hyperparameters. Please see details in Appendix A.

5.2 Evaluation Metrics

We employ three complementary metrics to jointly evaluate reasoning correctness and spatial grounding:

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

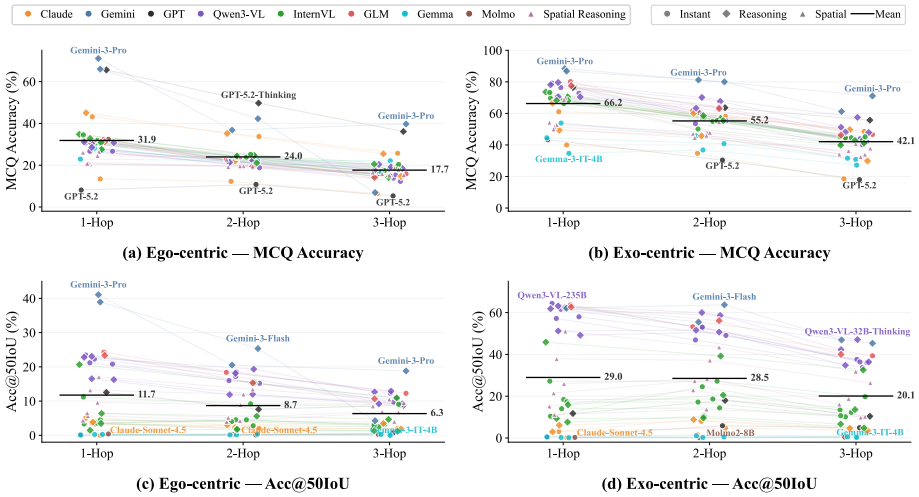


Fig. 5: All model performance across hop count and ego/exo perspectives.

Acc@50IoU. Our primary grounded metric requires correct 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.
Avg. IoU. Computed exclusively over MCQ-correct samples, this metric isolates grounding capability from reasoning errors, evaluating how precisely a model localizes the target once correctly identified.

5.3 Main Results

Overall Performance Highlights. As shown in Tab. 1, Gemini-3-Pro achieves the highest MCQ accuracy (64.7%) and Acc@50IoU (40.6%), while Qwen3-VL-32B-Thinking leads the open-weight category. Notably, top performance in answer selection does not guarantee precise localization (*e.g.*, Qwen3-VL-235B leads in average IoU). This discrepancy demonstrates that compositional reasoning and grounding capabilities remain largely decoupled in current VLMs.
Metric-dependent Rankings. This decoupling causes drastic rank reversals between metrics. Models like Claude-Opus-4.5 and Molmo2-8B rank high in MCQ but plummet in Acc@50IoU (*e.g.*, Claude drops from 7th to 29th), indicating shortcut-based answers without genuine localization. Conversely, the smaller Qwen3-VL-4B rises from 25th to 10th due to its robust grounding. These shifts prove that evaluating MCQ alone is highly misleading, making Acc@50IoU essential for verifying 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 (8.5%) and Claude-Sonnet-4.5-Thinking (1.9%) fail drastically here, confirming our benchmark rigorously evaluates both compositional reasoning and spatial grounding capabilities of current VLMs.

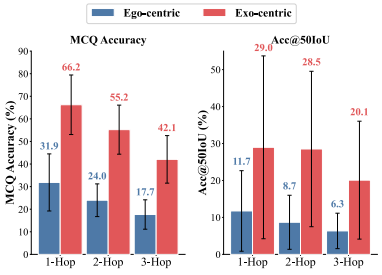


Fig. 6: Average Performance by Hop Count.

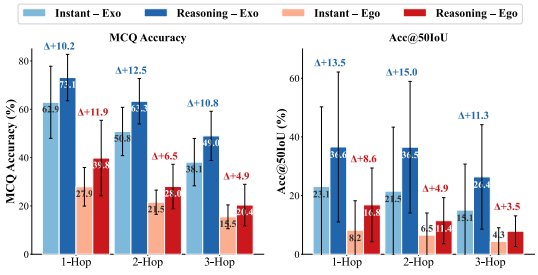


Fig. 7: Reasoning vs. Instant Models.

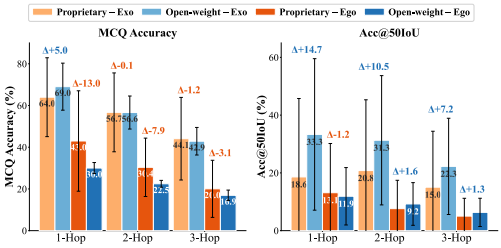


Fig. 8: Open-weight vs. Proprietary models.

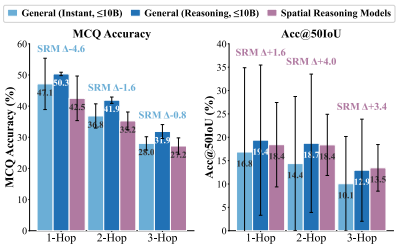


Fig. 9: Spatial Reasoning Models vs. General Models (≤10B).

5.4 In-depth Analysis

In this section, we present a comprehensive analysis of 37 VLMs on the proposed MultihopSpatial benchmark and derive eight key insights.

A1. Performance Degradation across Reasoning Hops. Evaluating the impact of reasoning complexity across all 37 models (Fig. 6) reveals a consistent performance degradation as the number of hops increases, confirming that compositional spatial reasoning remains a fundamental challenge for current VLMs. Both MCQ accuracy and Acc@50IoU exhibit steep declines from 1-hop to 3-hop. Crucially, this degradation is exacerbated under ego-centric evaluation, which requires additional perspective transformation. The widening performance gap between ego-centric and exo-centric views at higher hops suggests that perspective-taking compounds with multi-step reasoning, creating a multiplicative rather than additive difficulty.

A2. Ego vs. Exo: Perspective-Taking as a Compounding Bottleneck. Beyond a simple performance drop (Fig. 6), ego-centric evaluation fundamentally alters capability visibility. Under exo-centric conditions (Fig. 5(d)), our Acc@50IoU metric clearly distinguishes grounding-capable models (*e.g.*, Qwen3-VL) from those lacking native localization (*e.g.*, InternVL-3.5) despite their similar MCQ accuracies. Conversely, ego-centric evaluation (Fig. 5(c)) acts as an evaluation blind spot; it suppresses even strong grounding models to a 20–25% floor, completely masking these capability gaps. This compression highlights the necessity of evaluating perspectives jointly and reinforces Acc@50IoU as an essential metric to expose disparities invisible to conventional MCQ evaluation.

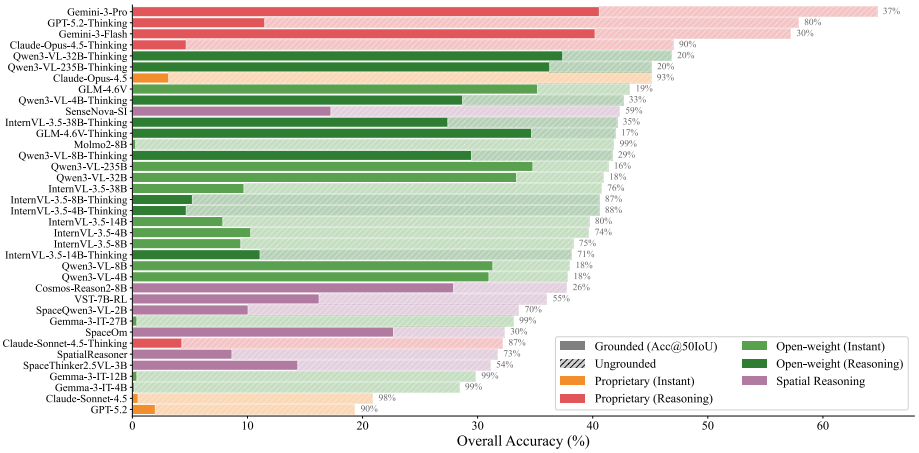


Fig. 10: Grounded (Acc@50IoU) vs. Ungrounded accuracy per model. Solid bars indicate grounded accuracy (Acc@50IoU) and hatched regions indicate ungrounded responses. Percentages denote the ungrounded ratio.

A3. Instant vs. Reasoning: Diminishing Returns under Multi-Hop Pressure. While reasoning models consistently outperform instant models (Fig. 7), this advantage diminishes under multi-hop pressure. The initial MCQ gain of up to +8 pp at 1-hop narrows considerably by 3-hop. Critically, even with extended thinking, reasoning models plummet to sub-20% MCQ accuracy and sub-10% Acc@50IoU on 3-hop ego-centric tasks. This convergence toward the performance floor reveals that inference-time reasoning (*e.g.*, chain-of-thought) yields diminishing returns as compositional steps accumulate. MultihopSpatial thus exposes a capability ceiling that test-time compute alone cannot resolve, confirming its value as a genuinely challenging benchmark even for state-of-the-art models.

A4. Open-weight vs. Proprietary Models. While proprietary models hold a slight MCQ advantage in ego-centric tasks (Fig. 8)—indicating stronger perspective-taking—open-weight models consistently dominate in Acc@50IoU. This reversal stems from a distinct asymmetry in visual grounding capabilities. Proprietary models exhibit high variance: while the Gemini-3 series demonstrates robust localization, others (*e.g.*, GPT-5.2, Claude) lack native grounding, significantly dragging down the group average. Conversely, open-weight families (*e.g.*, Qwen3-VL, GLM-4.6V) maintain consistently high grounding performance, forming a dense cluster at 60%+ in exo-centric Acc@50IoU. These findings highlight that the open-weight ecosystem currently offers more reliable spatial grounding than its proprietary counterparts.

A5. Generalist vs. Specialist Models. We compare specialized spatial reasoning models (SRMs) with general-purpose VLMs of comparable scale ($\leq$10B) in Fig. 9. Despite being explicitly tailored for spatial reasoning, SRMs are consistently outperformed by generalists on multi-hop MCQ accuracy at every hop count, although the gap narrows monotonically as the reasoning chain deepens ($-4.6 \rightarrow -1.6 \rightarrow -0.8$ from 1- to 3-hop). This indicates that current spatial spe-

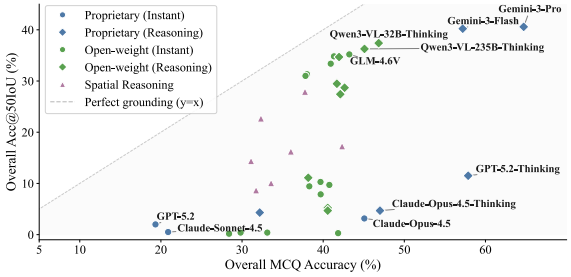


Fig. 11: Grounding Gap.

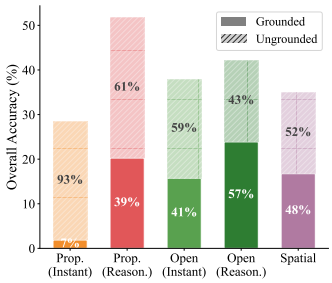


Fig. 12: Ungrounded Ratio.

cialization does not transfer to the compositional, multi-hop reasoning that our benchmark targets. The ordering only reverses under Acc@50IoU, which counts a prediction as correct only when the target is additionally localized (IoU $\geq 50\%$): here SRMs surpass their instant generalist counterparts at every hop, and the margin widens with hop depth (Δ+1.6, +4.0, +3.4). Since this metric further requires accurate localization, the reversal indicates that the advantage of SRMs is confined to grounding rather than reasoning—even with lower MCQ accuracy, they localize correctly answered targets more reliably. Generalists are thus stronger at answer selection while SRMs are stronger at localization, and these complementary strengths suggest that robust multi-hop spatial understanding requires the joint optimization of reasoning and grounding, which neither family currently attains.

A6. Grounding gap: Grounded vs. Ungrounded Accuracy. Fig. 11 reveals a striking disconnect between answer selection and spatial localization: most models fall far below the $y = x$ line, indicating that high MCQ accuracy does not entail accurate grounding. Across all 37 models, the average ungrounded ratio reaches 59%, meaning that more than half of correctly answered questions lack proper spatial localization (Fig. 10). This ratio varies dramatically by category (Fig. 12): proprietary instant models exhibit 93% ungrounded ratio, while open-weight reasoning models achieve the lowest at 43%, largely driven by Qwen3-VL and GLM families whose ungrounded ratios remain below 20%. At the extreme, models such as Gemma-3-IT, Molmo2, and Claude-Sonnet-4.5 exceed 98% ungrounded, effectively answering through shortcuts without any spatial understanding. These findings validate Acc@50IoU as an essential complement to MCQ evaluation: without it, models that rely entirely on linguistic shortcuts would be indistinguishable from those with genuine spatial understanding.

A7. The Limits of LLM Scaling in Spatial Reasoning. Analyzing model scale across three open-weight families (Fig. 13) reveals that scaling the language backbone alone yields limited gains. While MCQ accuracy shows modest, saturating improvements, Acc@50IoU largely plateaus (*e.g.*, Qwen3-VL) or remains near zero (*e.g.*, Gemma-3-IT). A notable exception is InternVL-3.5 at 38B, which exhibits a sharp grounding jump from 5.2% to 27.4%. Crucially, this leap coincides with a significant upgrade in its vision encoder (from 300M to 6B), whereas

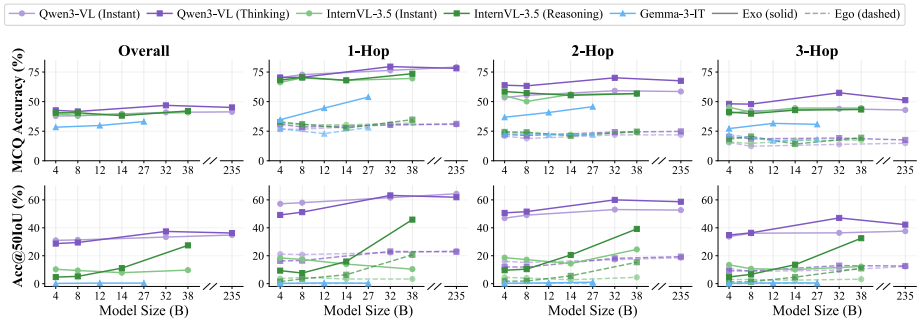


Fig. 13: Effect of model scale on MCQ Accuracy (top) and Acc@50IoU (bottom) across overall and per-hop settings. Solid lines indicate exo-centric and dashed lines indicate ego-centric performance.

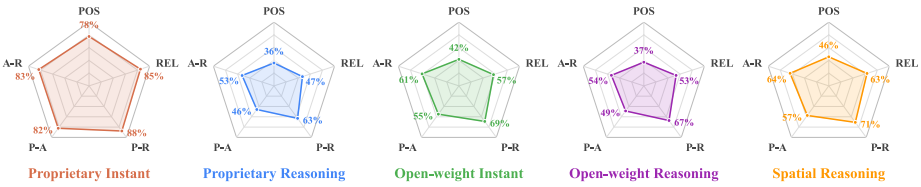


Fig. 14: Average Error Rates (↓, %) by tag combination across model categories. A, R, and P denote Attribute, Relation, and Position, respectively.

models utilizing fixed, small vision encoders (*e.g.*, Qwen3-VL's 400M) saturate early. This stark contrast demonstrates that multi-hop spatial reasoning depends critically on scaling the capacity of visual spatial representations, rather than relying solely on the language component's reasoning capacity.

A8. Error Analysis on Tag Compositions. Analyzing error rates across spatial tag combinations (Fig. 14) reveals that while reasoning models consistently outperform instant models, multi-tag compositions remain a critical bottleneck. Specifically, the POS-REL (P-R) combination incurs significantly higher errors than single-tag settings, highlighting the severe difficulty of jointly handling positional localization and relational comparison. Even specialized spatial reasoning models, which achieve the lowest overall errors, exhibit non-trivial failures on these complex compositions. This underscores that true compositional spatial reasoning remains an unresolved challenge across the current VLM ecosystem.

5.5 MultihopSpatial-Train: Utility as a Training Corpus

To validate MultihopSpatial as a training resource for improving spatial reasoning in VLMs, we compare our Qwen3-VL-4B-Instruct trained on our MultihopTrain with its baseline on spatial reasoning benchmarks and VLA benchmarks.

Spatial Reasoning Benchmarks. We evaluate the GRPO-trained model on both the in-domain MultihopSpatial benchmark and five out-of-domain spatial reasoning benchmarks in Tab. 2. On the in-domain MultihopSpatial benchmark,

Table 2: Impact of MultihopSpatial-Train on spatial reasoning across diverse benchmarks.

Model	MultihopSpatial			Out-of-Domain				
	Acc.	Acc@50IoU	avg. IoU	BLINK [14]	3DSRBench [26]	OmniSpatial [19]	VSI-Bench [40]	SpatialMQA [24]
Qwen3-VL-4B-Instruct	37.8	31.0	69.9	82.5	56.1	42.7	62.8	39.6
w/ MultihopSpatial-Train	**62.9**	**53.8**	**72.6**	**85.3**	**56.3**	**43.9**	**63.2**	**41.1**

Table 3: Vision-Language-Action evaluation on Calvin ABC-D [28] and Libero [23] benchmark. Each task (Task-#) measures success rate (%) on Calvin and Calvin denotes the average number of successfully completed tasks per sequence.

VLM backbone	Task-1	Task-2	Task-3	Task-4	Task-5	Calvin	Libero
Qwen3-VL-4B-Instruct	92.4	81.8	74.1	66.8	59.9	3.75	35.8
w/ MultihopSpatial-Train	**93.0**	**85.4**	**79.3**	**73.2**	**66.9**	**3.98**	**40.0**

GRPO training yields substantial improvements across all three metrics. The large gains in MCQ Accuracy and Acc@50IoU indicate that the model learns to perform multi-hop spatial reasoning more reliably, while the modest but consistent improvement in Avg. IoU suggests that grounding quality also benefits. More notably, the trained model achieves consistent improvements on all five out-of-domain benchmarks, despite being trained exclusively on MultihopSpatial data. This suggests that the spatial reasoning capabilities acquired through our dataset generalize beyond the in-domain distribution, transferring to diverse spatial understanding tasks. Additional experimental results with different model scales are described in Sec. B.4.

VLM as backbone for Vision-Language-Action. To investigate whether improved spatial reasoning transfers to embodied decision-making, we evaluate the models as VLM backbones using VLM4VLA [43] on the CALVIN ABC-D [28] and Libero [23] benchmarks. As shown in Tab. 3, the MultihopSpatial-trained model consistently outperforms the baseline. On CALVIN, the average task completion score rises from 3.75 to 3.98, with the performance gap widening as the task chain lengthens (from +0.6% on Task-1 to +7.0% on Task-5). This highlights compounding benefits in long-horizon sequential manipulation. Furthermore, the model achieves a +4.2% absolute improvement (35.8% → 40.0%) on the Libero benchmark. These results demonstrate that spatial reasoning capabilities acquired from MultihopSpatial-Train not only generalize across vision-language benchmarks but also translate into robust robotic policy execution, reinforcing spatial reasoning as a foundational capability for VLAs.

5.6 Qualitative Results

Fig. 15 illustrates a representative failure case for a 3-hop ego-centric question, requiring three sequential inferences. The query asks to identify the farthest square or rectangular object *in front of* the sitting person. However, all three models neglect POS, instead selecting the farthest rectangular object based solely on ATT and REL. Examining each model's rationale reveals that they explicitly

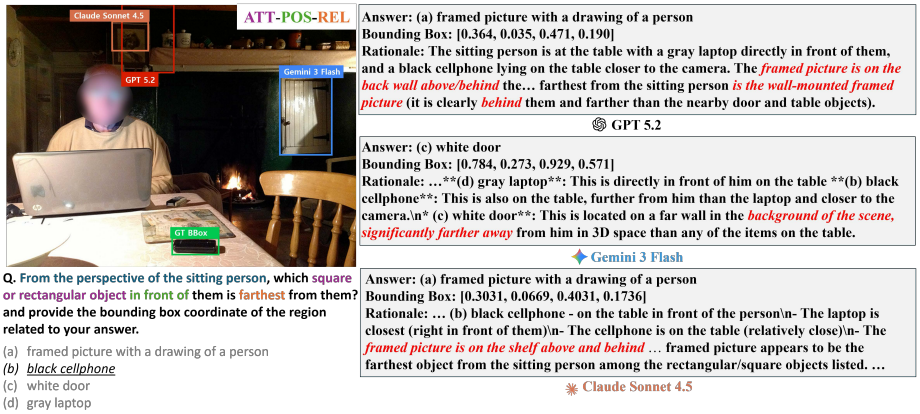


Fig. 15: Qualitative failure analysis on a 3-hop ego-centric question. All three models recognized the position condition *"in front of"*, but failed to incorporate it into their final predictions.

acknowledge the *in front of* constraint during reasoning, yet consistently disregard it in the final answer—demonstrating a failure to maintain intermediate conditions throughout the chain. This highlights the necessity of our compositional evaluation: an error at any single hop inevitably propagates to an incorrect final prediction. Additional qualitative results are provided in Appendix F.

6 Conclusion

We introduce **MultihopSpatial**, a comprehensive dataset for evaluating and enhancing multi-hop, compositional spatial reasoning in VLMs. To ensure evaluations explicitly reflect the visual grounding vital for Vision-Language-Action (VLA) deployment, we propose **Acc@50IoU**, a complementary grounded metric that simultaneously evaluates correct answer selection and precise spatial localization. While our extensive evaluation of 37 models reveals that complex spatial reasoning remains a formidable challenge, we demonstrate that RL post-training on our dedicated corpus effectively mitigates this issue. Ultimately, MultihopSpatial drives consistent improvements in both intrinsic VLM reasoning and downstream VLA task execution. We hope our dataset and grounded evaluation paradigm will catalyze future research in Embodied AI, particularly in scaling up spatial priors for real-time, dynamic robotic control in the wild.

7 Acknowledgments

This work was supported by 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, 90%) and (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST), 10%).

References

1. AI, R.: Spaceom. https://huggingface.co/remyxai/SpaceOm (2024), accessed: 2026-03-02
2. AI, R.: Spacethinker-qwen2.5vl-3b. https://huggingface.co/remyxai/SpaceThinker-Qwen2.5VL-3B (2025), accessed: 2026-03-02
3. Alam, N., Murali, L.K., Bharadwaj, S., Liu, P., Chung, T., Sharma, D., Kiran, K., Tam, W., Vegesna, B.K.S., et al.: The spatial blindspot of vision-language models. arXiv preprint arXiv:2601.09954 (2026)
4. Anthropic: Claude opus 4.5 system card. Tech. rep., Anthropic (2025), https://www.anthropic.com/claude-opus-4-5-system-card, accessed: 2026-03-02
5. Anthropic: Claude sonnet 4.5 system card. Tech. rep., Anthropic (September 2025), https://www.anthropic.com/claude-sonnet-4-5-system-card, accessed: 2026-03-02
6. 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)
7. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL. PMLR (2023)
8. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., et al.: Scaling spatial intelligence with multimodal foundation models. arXiv preprint arXiv:2511.13719 (2025)
9. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: CVPR (2024)
10. Chen, H., Guo, J., Wang, B., Zhang, T., Huang, X., Zheng, B., Hou, Y., Tie, C., Deng, J., Shao, L.: Goal-vla: Image-generative vlms as object-centric world models empowering zero-shot robot manipulation. arXiv preprint arXiv:2506.23919 (2025)
11. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
12. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
13. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 346–355 (2024)
14. 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. In: ECCV. pp. 148–166. Springer (2024)
15. GLM-V Team, Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., et al.: GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025), https://arxiv.org/abs/2507.01006
16. Google: Gemini 3. https://gemini.google.com/ (2025), accessed: 2026-03-02
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)

18. Huang, H., Chen, X., Chen, Y., Li, H., Han, X., Wang, Z., Wang, T., Pang, J., Zhao, Z.: Roboground: Robotic manipulation with grounded vision-language priors. In: CVPR (2025)
19. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In: ICLR (2026), https://openreview.net/forum?id=6nZKT2rL0H
20. Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
21. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. Springer (2014)
23. Liu, B., Zhu, Y., Gao, C., Feng, Y., qiang liu, Zhu, Y., Stone, P.: LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In: NeurIPS (2023), https://openreview.net/forum?id=xzEtNSuDJk
24. 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: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) ACL. pp. 620–632. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.31, https://aclanthology.org/2025.acl-long.31/
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018), https://openreview.net
26. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In: ICCV. pp. 6924–6934 (2025)
27. Ma, W., Chou, Y.C., Liu, Q., Wang, X., de Melo, C., Xie, J., Yuille, A.: Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In: NeurIPS (2025)
28. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters **7**(3), 7327–7334 (2022)
29. NVIDIA: Cosmos-reason2-8b. https://huggingface.co/nvidia/Cosmos-Reason2-8B (2025), accessed: 2026-03-02
30. OpenAI: Update to GPT-5 system card: GPT-5.2. Tech. rep., OpenAI (December 2025), https://openai.com/index/gpt-5-system-card-update-gpt-5-2/, accessed: 2026-03-02
31. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: CVPR (2023)
32. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR (2019)
33. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Li, X., Zhang, M., Zhang, Y., Li, Y., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
34. 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)

35. Wang, Z., Li, J., Zheng, J., Zhang, W., Liu, D., Zheng, Y., Niu, H., Yu, J., Zhan, X.: Physiagent: An embodied agent framework in physical world. arXiv preprint arXiv:2509.24524 (2025)
36. Wasi, A.T., Faisal, W., Rahman, A., Anik, M.A., Shahriar, M., Topu, M.M., Meem, S.T., Priti, R.N., Mitu, S.A., Hoque, M.I., et al.: Spatialab: Can vision-language models perform spatial reasoning in the wild? arXiv preprint arXiv:2602.03916 (2026)
37. von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., Gallouédec, Q.: TRL: Transformers Reinforcement Learning. https://github.com/huggingface/trl (2020), apache-2.0 License
38. Wu, H., Huang, X., Chen, Y., Zhang, Y., Wang, Y., Xie, W.: Spatialscore: Towards comprehensive evaluation for spatial intelligence. In: CVPR (2026)
39. Yang, G., Zhang, T., Hao, H., Wang, W., Liu, Y., Wang, D., Chen, G., Cai, Z., Chen, J., Su, W., et al.: Vlaser: Vision-language-action model with synergistic embodied reasoning. arXiv preprint arXiv:2510.11027 (2025)
40. 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: CVPR. pp. 10632–10643 (2025)
41. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
42. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764 (2025)
43. Zhang, J., Chen, X., Guo, Y., Hu, Y., Chen, J.: VLM4VLA: Revisiting vision-language-models in vision-language-action models. In: ICLR (2026), https://openreview.net/forum?id=tc2UsBeODW
44. Zhang, Z., Li, H., Dai, Y., Zhu, Z., Zhou, L., Liu, C., Wang, D., Tay, F.E.H., Chen, S., Liu, Z., Liu, Y., Li, X., Zhou, P.: From spatial to actions: Grounding vision-language-action model in spatial foundation priors. In: ICLR (2026), https://openreview.net/forum?id=fzmittHfq3