OMNIVIEW-SPACE: REINFORCING SPATIAL REASONING VIA MULTI-PERSPECTIVE SPATIAL MAPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial intelligence remains a challenge for Multimodal Large Language Models (MLLMs), as it requires coherent spatial scene representations beyond basic object recognition. Existing methods typically build such representations through textual reasoning or 3D reconstruction. However, they often falter during multi-step reasoning, particularly when required to dynamically re-anchor evidence to the specific camera-, object-, or direction-centric reference frames demanded by complex queries. To address this, we propose **OmniView-Space**, a framework to maintain spatial consistency through multimodal egocentric evidence. Our approach consists of three core components: (1) Multi-Perspective Spatial Mapping (MPSM), which re-anchors reconstructed geometry into a query-aligned visual cognitive map and a textual spatial graph; (2) Tool-Guided Egocentric Reasoning, an interleaved policy trained to actively select the ego anchor required by the query and request the corresponding MPSM evidence; and (3) Cognitive-Map Distillation, which uses MPSM-generated trajectories and ego-frame rewards to train the model to reason with self-generated cognitive maps. Experiments on single- and multi-image spatial reasoning benchmarks show that OmniView-Space achieves state-of-the-art performance. Furthermore, the distilled model maintains this performance while reducing reliance on external geometry pipelines.

1 INTRODUCTION

Spatial intelligence refers to the ability to reason about the geometric and physical structure of the world from visual observations. Although Multi-modal Large Language Models (MLLMs) have made substantial progress in general visual understanding (Li et al., 2024; Chen et al., 2024b; Bai et al., 2025b;a), they still struggle with spatial questions involving relative localization, viewpoint taking, cross-view association, or camera-motion reasoning (Zhang et al., 2026b; Yang et al., 2025c; Zhang et al., 2026a). Unlike conventional visual question answering, which mainly focuses on semantic recognition in static images, spatial intelligence requires maintaining a consistent spatial state throughout the reasoning process. The challenge is therefore not only to recognize scene content, but also to preserve spatial consistency as intermediate reasoning steps unfold. For example, when asked what lies to the right of a chair from the point of view of the chair, knowing the room layout is insufficient. The model also needs to establish a chair-centered spatial state and reason within that reference frame; otherwise, judgments regarding left and right can easily become confused.

While textual chain-of-thought excels in symbolic and mathematical domains (Shao et al., 2024; Ouyang et al., 2025), it is less reliable for spatial intelligence (Cai et al., 2025b; Yang et al., 2025b). Purely textual traces often distort object correspondence, relative orientation, and topology, causing spatial errors to compound. To provide MLLMs with stronger spatial grounding, recent work explores two main forms of intermediate evidence (Fig. 1(a,b)). One form injects 3D visual evidence through 3D-aware inputs or external reconstruction (Luo et al., 2026; Zhang et al., 2026c). It preserves metric structure, but raw 3D outputs are often challenging for MLLMs to interpret. The other form constructs textual spatial graphs or grid maps (Yang et al., 2025a; Huang et al., 2025; Yin et al., 2025). It makes relations more language-accessible, but coordinate-style reasoning can lose fine-grained geometry and topology. More importantly, both forms usually express evidence in a system-chosen reference frame (e.g., the first camera view or a world system), creating a mismatch with many spatial questions that require a *query-specified ego frame*. As shown in Fig. 1, individual questions anchor this frame to distinct *ego anchors*, such as cameras, objects, or directions. How-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

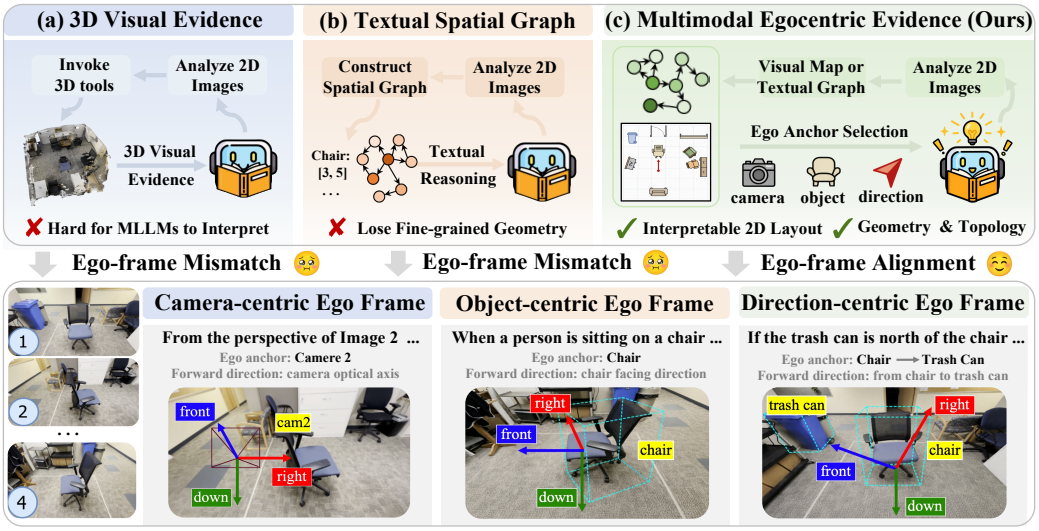

Figure 1: (a) 3D visual evidence preserves the metric structure, but raw reconstruction outputs are difficult for MLLMs to interpret. (b) Textual spatial graphs are language-accessible, but coordinate-style reasoning can lead to the loss of fine-grained geometry. (c) OmniView-Space builds multi-modal egocentric evidence: a visual BEV cognitive map and a textual spatial graph aligned to the query-specified ego frame. Bottom: existing evidence can be frame-mismatched, while OmniView-Space aligns it to the camera-, object-, or direction-centric frame required by the question.

ever, existing methods rely on textual coordinate representations to translate system-frame evidence into the query-specified frame, where performing such a geometric transformation purely through text is highly unreliable and easily breaks spatial consistency during multi-step reasoning.

Humans handle such ego-frame shifts naturally: we can keep a coherent mental scene, imagine it from the query-specified ego frame, and then make a spatial judgment. In contrast, current MLLMs often struggle when the query-specified ego frame differs from the observed camera or system-defined frame (Ma et al., 2025a; Zhang et al., 2024). To address this challenge, we propose **OmniView-Space**, a framework for multimodal egocentric evidence construction and reasoning, as shown in Fig. 1(c). This design reframes diverse camera-, object-, and direction-centric queries as reasoning in the query-specified ego frame, following the intuition that if a model can perceive the scene in this frame, spatial reasoning becomes easier to ground. The framework uses two complementary forms of MPSM evidence: a *visual BEV cognitive map* for inspectable 2D layout and a *textual spatial graph* for explicit relational structure. It follows a progressive design that moves from external spatial evidence to tool-using reasoning and finally to self-generated cognitive-map reasoning. ① First, we develop *Multi-Perspective Spatial Mapping* (MPSM), which constructs multimodal egocentric evidence from multi-view images, task-relevant objects, and the ego anchor. MPSM returns both forms of evidence in the same query-specified ego frame, making object layout, relative direction, and topology directly usable for reasoning. ② Second, we train an interleaved tool-using policy with reinforcement learning (RL). The model learns to analyze the question, identify the ego anchor, request either the visual map or textual graph, and ground its answer in the returned multimodal egocentric evidence. ③ Finally, we distill this tool-guided behavior into the model itself. MPSM-generated trajectories teach the model to predict an egocentric cognitive map before answering, and during GRPO (Shao et al., 2024), MPSM further serves as a reward reference for refining the predicted objects, camera layout, and final answer.

Experiments on multi-image and single-image spatial reasoning benchmarks show that OmniView-Space significantly improves spatial reasoning over strong MLLM baselines. Notably, the distilled model closes much of the gap to the tool-augmented pipeline while removing the need to invoke the full geometry pipeline at inference time. Our contributions are summarized as follows:

- ❶ *Query-aligned egocentric evidence.* We identify reference-frame mismatch as a bottleneck in multi-step spatial reasoning: evidence is expressed in a system-defined frame, while spatial questions require judgment from a camera-, object-, or direction-centric ego frame. This motivates query-aligned egocentric evidence as the interface for spatial reasoning.

❷ *Multi-perspective spatial mapping.* We introduce the MPSM toolkit, which converts scene observations into multimodal egocentric evidence: a query-aligned visual BEV cognitive map and a textual spatial graph that combine 2D interpretability with explicit geometry.

❸ *Tool-guided reasoning and distillation.* We train an interleaved tool-using policy with RL to select the ego anchor required by the query and actively call MPSM for visual cognitive map or textual spatial graph construction. We then distill this tool-guided reasoning into self-generated cognitive maps using MPSM-generated trajectories and ego-frame rewards.

2 Related Work

Spatial reasoning in MLLMs. Spatial reasoning remains a foundational challenge for Multimodal Large Language Models (MLLMs) (Li et al., 2026; Yang et al., 2025b). Recent progress has been driven by stronger MLLMs (Bai et al., 2025a; Li et al., 2024), broader benchmarks (Yang et al., 2025c;a), and training or prompting strategies that use 3D reconstruction, depth cues (Roy et al., 2025), large-scale 3D spatial VQA data (Balazadeh et al., 2024; Zhang et al., 2025), spatial prompts (Liu et al., 2025a; Marsili et al., 2025), mental simulation (Chen et al., 2025b; Yin et al., 2025), visual chain-of-thought (Fan et al., 2025; Wang et al., 2025c; Lee et al., 2025), and explicit visual grounding (Wu et al., 2026). In multi-view settings, multiple observations can help recover 3D structure and reduce single-view ambiguity, and multiview-augmented MLLMs (Daxberger et al., 2025; Chang et al., 2025; Lee et al., 2025; Wu et al., 2025; Xu et al., 2025a; Zheng et al., 2025a) improve geometric understanding and perspective taking. However, these advances still leave a gap in maintaining spatial evidence in the query-specified ego frame, especially when the answer must be judged from a camera-, object-, or direction-centric perspective rather than a system-defined frame.

Geometry-augmented tool use. Tool calling strengthens MLLMs by invoking external specialists through prompting, code generation, or learned policies (Chen et al., 2025a; Zhang et al., 2026c; Luo et al., 2026; Wu et al., 2026; Han et al., 2025). In spatial domains, these tools provide depth, segmentation, 3D reconstruction, camera poses, or viewpoint exploration as intermediate evidence (Lin et al., 2025; Carion et al., 2025; Wang et al., 2025a; Zhang et al., 2026c). Other approaches inject 3D-aware features or camera-guided representations directly into the network (Chen et al., 2025b; Zhao et al., 2025; Chen et al., 2026a). While demonstrating the value of explicit geometry, raw 3D outputs are often hard for MLLMs to interpret, widening the gap between geometry and language-level reasoning. Furthermore, their evidence remains fixed to a system-defined frame, motivating our approach to treat geometry as callable evidence aligned with the query's ego anchor.

Cognitive maps and frame grounding. Spatial cognition includes mental rotation, spatial visualization, and object assembly (Wang et al., 2025b; Xu et al., 2025b; Zha et al., 2025), and is often linked to Spatial Mental Models (SMMs) (Johnson-Laird, 1980; 1983). Recent MLLM studies probe these abilities via coordinate-aware prompting, CoT reasoning, representation alignment, and RL (Cai et al., 2025a; Liu et al., 2025b; Chen et al., 2024a; Cheng et al., 2024; Pan & Liu, 2025). A related cognitive-map paradigm uses coordinate or grid representations for spatial reasoning (Yang et al., 2025a; Chen et al., 2026b; Batra et al., 2025; Ouyang et al., 2025; Huang et al., 2025). However, many maps are discrete, fixed to a world or first-camera frame, or limited to single-frame inputs (Chang et al., 2025; Ma et al., 2025b), restricting their utility when answers depend on a query-specified ego frame. Recent allocentric-reasoning work similarly notes that MLLMs struggle when a query requires instantiating a target frame rather than using the observed view (Wang et al., 2026a). OmniView-Space addresses this by constructing multimodal egocentric evidence—including a visual BEV cognitive map and a textual spatial graph—and internalizing tool-guided reasoning through MPSM trajectories and ego-frame rewards.

3 Methodology

Given a question q and images $\mathcal{I}$, the proposed OmniView-Space grounds spatial reasoning in evidence expressed in the ego frame required by the query. We first define MPSM as an external evidence constructor (Sec. 3.1). As shown in Fig. 2, the model then analyzes the question and images, selects the ego anchor required by the query, calls MPSM for a query-aligned visual BEV cognitive map or textual spatial graph, and continues reasoning over the returned evidence (Sec. 3.2). Finally, we distill this tool-guided process into the model itself, so that it can reason with self-generated query-aligned cognitive maps without calling the full geometry pipeline at inference (Sec. 3.3).

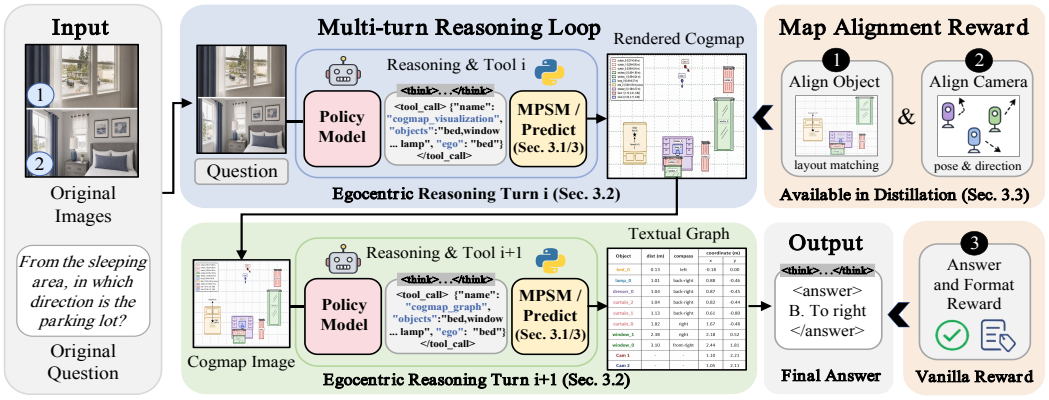

Figure 2: **Overview of the OmniView-Space.** Given a question and an image, the policy analyzes the ego anchor to either invoke MPSM (Sec. 3.1) or directly predict (Sec. 3.3) a query-aligned visual BEV cognitive map or textual spatial graph across multiple reasoning turns, where this trajectory is optimized via reinforcement learning using an answer reward (Sec. 3.2). During the distillation phase, MPSM generates trajectories for cold-start training and provides a reward for map alignment during RL, which instructs the model to reason using internally generated cognitive maps (Sec. 3.3).

3.1 Multi-Perspective Spatial Mapping

Role and interface. As shown in Fig. 3, MPSM is an evidence constructor that takes the images $\mathcal{I}$ as input, along with task-relevant object hints $\mathcal{H}$ and an ego anchor r **predicted by the MLLM**. The hints $\mathcal{H}$ can be object names and 2D boxes, while r can be an input camera, an object with an estimated facing direction, or a direction defined by a source-to-target object relation. The MPSM operator Ω returns two forms of the same query-aligned scene evidence: `cogmap_visualization` renders a visual BEV cognitive map Π_r, while `cogmap_graph` exports a textual spatial graph $\mathcal{T}_r$ with object attributes, containment/support edges, distances, and ego-relative directions. Formally,

$$\mathcal{G}_r = \Omega(\mathcal{I}, \mathcal{H}, r) := \{r, \Pi_r, \mathcal{T}_r\}, \tag{1}$$

where $\mathcal{G}_r$ is the spatial state re-anchored to ego reference r, Π_r and $\mathcal{T}_r$ are the rendered BEV cogmap and the relation graph expressed in the same ego frame. Here, the index r denotes outputs transformed into the ego frame (Eq. 3). To construct this spatial state, MPSM performs two main steps:

Step 1: Default-Frame Scene Reconstruction. To reconstruct the 3D scene from the input views, MPSM leverages a dense geometry model (Lin et al., 2025) that predicts the camera pose alongside a per-pixel point map X_t, validity mask U_t, and confidence map C_t for each image $I_t \in \mathbb{R}^{H\times W}$. Concurrently, a promptable segmenter (Carion et al., 2025) grounds the object hints $\mathcal{H}$ into foreground masks using text prompts or bounding boxes. We retain only the pixels and frames that pass a confidence threshold,

$$V_t = \mathbf{1}[U_t{=}1] \cdot \mathbf{1}[C_t{>}\tau_{\text{pix}}] \cdot \mathbf{1}[\bar{C}_t{>}\tau_{\text{frm}}],$$
$$\bar{C}_t = \frac{1}{HW}\sum_{u,v} C_t(u,v). \tag{2}$$

where V_t is the retained-pixel mask, $\bar{C}_t$ is the average frame confidence, and τ_{pix} and τ_{frm} are pixel- and frame-level thresholds. Retained foreground points are fused across views, voxelized, assigned semantic labels by majority vote, and clustered into object instances. Each instance is represented by a 3D box and BEV footprint, with its facing direction estimated using (Wang et al., 2026b) when required for object-centric queries.

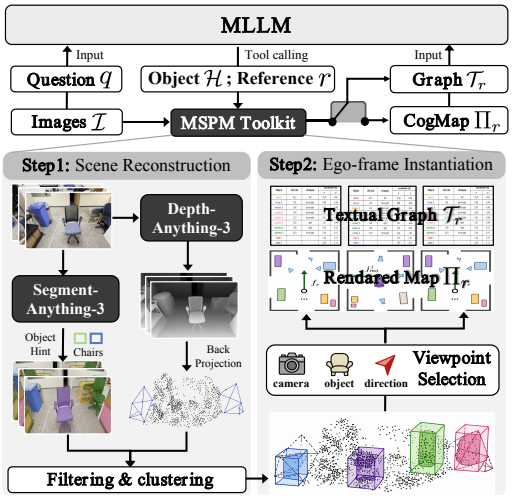

Figure 3: **MPSM toolkit.** Given multi-view images, object hints, and an ego reference, MPSM reconstructs a metric scene, re-anchors it to the query-specified frame, and returns both a rendered BEV cogmap and a textual spatial graph.

Step 2: Ego-frame Instantiation. Since the reconstructed scene is initially expressed in the default reference frame, we should re-express it in the query-specified ego frame. After gravity alignment, the ego anchor r defines an origin o_r and a forward direction f_r on the ground plane. The pair (o_r, f_r) is instantiated differently for each anchor type: (1) a camera anchor uses the camera center and optical axis, (2) an object anchor uses the object centroid and estimated facing direction, and (3) a direction anchor uses a source-object centroid with the source-to-target vector as forward. Let $\tilde{p} \in \mathbb{R}^3$ be a point in the gravity-aligned scene coordinate system, and let R_r be the yaw rotation that maps f_r to the positive BEV Y axis. The coordinate of $\tilde{p}$ in the query-specified ego frame is:

$$p^{(r)} = R_r(\tilde{p} - o_r). \tag{3}$$

The same transformation is applied to object boxes, BEV footprints, and camera poses. MPSM renders Π_r with the ego at the origin and exports $\mathcal{T}_r$ with the same ego-relative geometry, so the visual BEV cognitive map and textual spatial graph remain consistent views of one spatial state.

3.2 Tool-Guided Egocentric Reasoning

To leverage MPSM's egocentric evidence, we train the MLLM as an interleaved tool-use policy. The model learns to analyze the query, select the required ego anchor and object hints, invoke the visual BEV cognitive map or textual spatial graph, and ground its answer in the returned spatial state.

Interleaved rollout. At the decoding step t, the policy conditions on both generated text and multimodal observations returned by tools. We denote the rollout state and action distribution as

$$s_t = \{X_{\le t};\, M_{\le t}\}, \qquad a_t \sim \pi_\theta(a \mid s_t), \tag{4}$$

where $X_{\le t}$ is the model-generated text history, $M_{\le t}$ contains the original images and any MPSM observations, a_t is the next text or tool action, and π_θ is the MLLM policy with parameters θ. Tokens from tool observations are masked from the policy gradient, so optimization is applied only to the model's analysis, tool parameters, reasoning, and answer.

Complementary visual and textual evidence. Prior spatial reasoning methods often keep intermediate maps in textual or coordinate-based form, requiring the model to infer spatial relations from serialized layouts or JSON-style representations (Yang et al., 2025a; Yin et al., 2025). We train and evaluate two variants on the MindCube benchmark, one using text-coordinate maps and the other using visual BEV cognitive maps. Both interfaces are generated from the same query-aligned state $\mathcal{G}_r$; they differ in how that state is presented to the MLLM. As shown in Fig. 4, the visual-map condition achieves better overall performance and elicits higher image-token attention (Appendix H), suggesting that preserving BEV layouts visually provides stronger grounding for qualitative viewpoint reasoning such as left–right, front–back, and between-object judgments. The textual spatial graph remains complementary: it provides more explicit numerical and structural evidence for distance comparisons, height or support relations, and object attributes, which is consistent with the tool-call analysis in Fig. 8, where graph calls become more frequent on 3D relation tasks. The original views remain reliable for visual identity. We therefore allow the policy to choose among the original images, the visual BEV cognitive map Π_r, and the textual spatial graph $\mathcal{T}_r$ according to the spatial demand of each query, rather than imposing a fixed modality schedule.

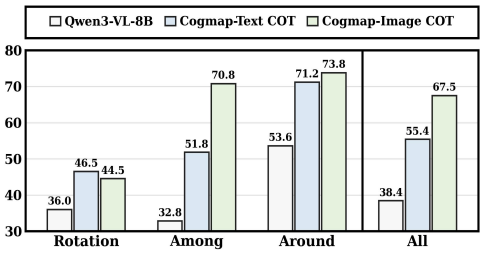

Figure 4: **Comparison of evidence formats.** On MindCube, using a visual BEV cognitive map outperforms text-coordinate reasoning overall, suggesting that visualized egocentric evidence is a more usable intermediate representation for current MLLMs.

Policy optimization. During a rollout, the model interactively invokes the MPSM tool—specifying the target hints, ego reference, and desired modality (visual map Π_r or relation graph $\mathcal{T}$)—to gather spatial observations. For a complete trajectory τ terminating with a final answer $\hat{a}$, we optimize the MLLM policy via GRPO (Shao et al., 2024) with the reward objective:

$$R_{\text{tool}}(\tau) = R_{\text{ans}}(\hat{a}) + \lambda_{\text{fmt}} R_{\text{fmt}}(\tau), \tag{5}$$

where R_{ans} scores answer correctness, R_{fmt} enforces the tool-call and answer format (Zheng et al., 2025b), and λ_{fmt}=0.2 controls the strength of the format reward. This trajectory-level objective trains ego-frame selection, tool invocation, map/graph reading, and final answering jointly.

3.3 Distilling Cognitive-Map Reasoning

The tool-using policy provides reliable egocentric evidence, but it still depends on external geometry tools. Next, we will explore how to distill the tool-guided reasoning ability into the model: given $(q, \mathcal{I})$, the model learns to predict a JSON-format ego-frame cognitive map, which is rendered into a visual BEV cognitive map for subsequent reasoning, as shown in Fig. 2.

MPSM-guided SFT Data. To enable the model to actively generate such cognitive maps without relying on external tools, we first collect expert trajectories that combine geometric tool outputs with reasoning traces. Specifically, we construct Supervised Fine-Tuning (SFT) trajectories, leveraging MPSM for intermediate cognitive mapping and Qwen3-VL-235B (Bai et al., 2025a) for generating reasoning traces. For each sample, the MLLM first extracts relevant object hints and predicts the required ego anchor. Next, MPSM constructs a query-aligned visual BEV cognitive map (Sec. 3.1), denoted as J_s. The MLLM then generates two rationales: R_1, which explains the construction of J_s, and R_2, which infers the predicted answer $\hat{a}_s$ from the inputs and J_s. The resulting SFT trajectory is assembled as: `<think>` R_1 `</think>` $\rightarrow$ `<cogmap>` J_s `</cogmap>` $\rightarrow$ [*Rendered Image*] $\rightarrow$ `<think>` R_2 `</think>` $\rightarrow$ `<answer>` $\hat{a}_s$ `</answer>`. Before training, we filter noisy trajectories via answer checking, critic-based repair, and missing-object refinement. Refer to the appendix for more details.

Map Alignment Reward. SFT initializes the model to generate cogmaps and map-grounded rationales, but 1D token-level imitation does not explicitly reward 2D map geometric alignment. We therefore further train with GRPO using both answer rewards and map-quality rewards. Let $\hat{J}$ be the predicted JSON-format cogmap and $\hat{a}$ be the final answer. The total reward is defined as:

$$R(\hat{J}, \hat{a}) = R_{\text{ans}}(\hat{a}) + \lambda_{\text{fmt}} R_{\text{fmt}}(\hat{J}, \hat{a}) + g(\hat{a}) \Big(\alpha \cdot R_{\text{object}}(\hat{J}) + \beta \cdot R_{\text{cam}}(\hat{J}) \Big), \tag{6}$$

where R_{ans} and R_{fmt} evaluate answer correctness and format compliance, α, β are hyper-parameters balancing different reward components, and $g(\hat{a}) = \mathbf{1}[\hat{a} = a^*]$ applies map alignment only to correct predictions. The alignment term has two parts. (1) **Object alignment** $R_{\text{object}} = \sum_{c \in \mathcal{C}} \frac{n_c}{N_o} R_c$, which first performs category-wise Hungarian matching in the query-specified ego BEV frame, then computes R_c from matched object pairs using BEV bbox center accuracy and size consistency. (2) **Camera alignment** $R_{\text{cam}} = \frac{1}{N_c} \sum_i (r^i_{\text{pos}} + r^i_{\text{ori}})$ measures camera position, pairwise direction consistency, and facing-angle agreement, with $r^i_{\text{ori}} = (1 + \cos \Delta\theta_i)/2$.

4 Experiments

4.1 Experimental Settings and Evaluation Benchmarks

Implementation Details. OmniView-Space is built on Qwen3-VL-4B (Bai et al., 2025a). For the supervised fine-tuning (SFT) stage, we utilize 26K training data (Refer to the Appendix B for details). For the reinforcement learning (RL) phase in both tool-guided reasoning and cognitive-map distillation, the training data is drawn from an 11K subset filtered from SpatialLadder-26K (Li et al., 2025) and an additional 10K samples from MindCube (Yin et al., 2025). The tool-guided reasoning stage is trained directly with GRPO, without SFT warm-up. Cognitive-map distillation uses both SFT and RL: we first perform SFT for 3 epochs with a learning rate of 1×10^{-5}, and then apply GRPO for 2 epochs with a smaller learning rate of 1×10^{-6} for stable policy updates. We sample $K = 8$ trajectories to compute group-relative advantages. All training is conducted on 8 GPUs, each equipped with 90GB of memory.

Evaluation Settings. We evaluate OmniView-Space on six spatial reasoning benchmarks covering multi-view aggregation, viewpoint-dependent reasoning, perspective transformation, and embodied spatial cognition. For multi-image reasoning, we report results on MindCube-Tiny (Yin et al., 2025), MMSI-Bench (Yang et al., 2025c), and SPAR-Bench (Zhang et al., 2026a). For single-image spatial reasoning, we evaluate on the Allocentric subset of 3DSRBench (Ma et al., 2025a), the Perspective subset of OmniSpatial (Jia et al., 2025), and SPBench Li et al. (2025). The results are reported as accuracy following (Yang et al., 2025a), and proprietary models are shown in gray for reference.

4.2 Main Results

Obs 1. Egocentric cognitive maps improve spatial reasoning across input settings. Table 1 shows that OmniView-Space outperforms all open-source baselines on three multi-image bench-

Table 1: **Main results on multi-view spatial reasoning benchmarks.** We report accuracy on MindCube-Tiny, MMSI-Bench, and SPAR-Bench. The **bold** and underlined indicate the best and second-best open-source results; proprietary models are shown in *gray italic* for reference.

Model	MindCube-Tiny				MMSI-Bench					SPAR-Bench				Overall
	Aro.	Rot.	Amo.	All	Pos.	Att.	Mot.	MSR	All	Low	Mid.	High	All	
Proprietary Models														
GPT-4.1 (Hurst et al., 2024)	*59.6*	*49.5*	*47.2*	*50.6*	*29.3*	*33.8*	*38.7*	*31.3*	*31.7*	*38.2*	*37.0*	*43.1*	*39.9*	*40.7*
GPT-5.2 (Singh et al., 2025)	*69.2*	*95.5*	*41.5*	*58.4*	*42.9*	*47.6*	*36.0*	*40.4*	*42.0*	*43.7*	*41.5*	*66.4*	*52.6*	*51.0*
Gemini-3-Pro (Pichai et al., 2025)	*51.6*	*54.5*	*46.0*	*49.0*	*42.9*	*50.0*	*38.0*	*40.9*	*42.7*	*40.6*	*21.7*	*58.0*	*43.1*	*44.9*
Open-source General Models														
InternVL3-8B (Zhu et al., 2025)	53.6	36.5	38.1	41.5	30.8	23.8	24.6	14.6	25.7	28.1	39.3	45.0	37.3	34.8
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	21.4	38.8	29.5	29.3	28.4	21.5	29.3	25.8	27.1	27.6	27.6	41.8	33.8	30.1
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	41.6	34.1	20.0	35.9	34.2	32.3	26.0	23.7	27.7	30.8	33.1	40.5	35.6	33.1
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	53.6	36.0	32.8	38.4	31.0	26.9	31.3	28.8	30.1	35.6	32.9	39.6	36.5	35.0
Spatial Intelligence Models														
MindCube-3B (Yin et al., 2025)	67.6	34.0	51.0	51.7	23.9	24.6	18.7	25.3	23.9	16.8	22.6	23.9	20.8	32.1
SpatialLadder-3B (Li et al., 2025)	50.8	35.0	43.2	43.4	30.3	23.3	16.0	21.2	25.4	24.9	29.4	42.4	32.9	33.9
Spatial-MLLM-4B (Wu et al., 2025)	36.0	39.0	30.5	33.4	28.5	25.4	18.0	26.3	26.1	22.3	32.4	35.2	30.4	30.0
SpaceR-7B (Ouyang et al., 2025)	49.2	35.0	34.2	37.9	29.1	29.4	21.9	22.5	26.9	27.0	23.9	44.6	34.2	33.0
ViLaSR-7B (Wu et al., 2026)	44.4	35.5	31.0	35.0	35.9	26.0	21.0	23.2	29.8	35.3	29.0	42.5	37.4	34.1
Cambrian-S-7B (Yang et al., 2026)	50.8	37.0	35.9	39.7	-	-	-	-	27.1	29.7	31.3	49.6	37.9	34.9
Tool-Integrated														
OmniView-Space-Zero	53.6	32.5	52.8	49.1	36.6	22.3	27.3	28.4	31.7	33.5	35.8	46.3	39.6	40.1
OmniView-Space-R1†	68.0	53.0	79.2	**71.5**	41.2	35.4	28.7	25.8	**35.5**	44.1	38.1	47.8	**44.8**	**50.6**
Tool-Distillation														
OmniView-Space-SFT	55.8	47.0	49.8	50.9	32.6	31.5	22.7	25.8	29.6	26.6	32.5	38.9	33.1	37.9
OmniView-Space-R1‡	71.2	48.5	69.2	65.7	36.2	31.5	24.7	24.7	31.6	43.2	34.3	44.2	42.3	46.5

Table 2: Results on the Allocentric subset of 3DSRBench. **Bold** and underlined numbers indicate the best and second-best performance.

Model	Front.	Left.	Twd.	Avg
RoboBrain2.0-32B (Team et al., 2025)	60.17	38.54	27.92	42.21
Spatial-SSRL-7B (Liu et al., 2025c)	57.41	33.95	21.72	37.69
Cosmos-Reason1-7B (Azzolini et al., 2025)	60.61	40.40	21.65	40.89
SenseNova-SI-1.1-7B (Cai et al., 2025b)	56.10	34.81	22.96	37.96
SpatialThinker-7B (Batra et al., 2025)	58.14	34.81	21.65	38.20
REVPT-7B (Zhou et al., 2025)	61.05	35.24	21.94	39.41
OmniView-Space-R1	**67.15**	**48.71**	**30.03**	**48.63**

Table 3: Results on the Perspective subset of OmniSpatial. **Bold** and underlined numbers indicate the best and second-best performance.

Model	Ego.	Allo.	Hypo.	Avg
SpaceMantis-13B (Chen et al., 2024a)	49.22	38.25	39.28	42.25
SpaceR-7B (Ouyang et al., 2025)	-	-	-	41.35
SpaceThinker-7B (Batra et al., 2025)	58.04	35.11	31.08	41.41
SpatialBot-3B (Cai et al., 2025a)	47.06	36.17	37.35	40.19
V2LO-7B (Huang et al., 2025)	-	-	-	46.70
RoboPoint-13B (Yuan et al., 2024)	49.02	37.66	33.49	40.06
OmniView-Space-R1	**63.73**	**42.02**	**57.83**	**54.53**

marks, scoring 71.5 on MindCube-Tiny, 35.5 on MMSI-Bench, and 44.8 on SPAR-Bench. Significant gains on MindCube-Tiny, which requires re-anchoring multi-view observations to query-specific viewpoints, suggest that egocentric cognitive maps provide a more effective reasoning interface than raw images or text alone. MMSI-Bench shows improved robustness across position, attribute, motion, and multi-step questions, while SPAR-Bench confirms the utility of explicit spatial evidence across difficulty levels. The tool-distilled model reaches 46.5, compared with 50.6 for the tool-integrated pipeline, suggesting that distillation retains part of the tool-guided gain. Beyond multi-image reasoning, Tables 2 and 3 evaluate single-image spatial reasoning, where models infer viewpoint-sensitive relations from one observation. OmniView-Space achieves the highest open-source average on the 3DSRBench allocentric tasks and leads on the Allo., Hypo., and Avg. metrics of OmniSpatial, while ranking second on Ego. These results indicate that egocentric map supervision improves internal spatial abstraction rather than only serving as a multi-view fusion mechanism.

Obs 2. Egocentric maps excel in re-anchoring and spatial relations. The fine-grained performance breakdown highlights the strengths of OmniView-Space. On MindCube-Tiny, the model excels in the Around, Rotation, and Among subtasks, which directly benefit from egocentric maps due to their reliance on complex relative positioning. On MMSI-Bench, it leads in Position and Attribute categories and remains competitive in Motion and Multi-step reasoning, indicating that explicit spatial grounding supports object-centric decisions. The SPAR-Bench results show similar utility across difficulty levels. Single-image results follow this trend: the model performs best on 3DSRBench (Front, Left, Towards) and OmniSpatial (Allo., Hypo.), which all demand scene transformation into query-specified frames. These consistent gains confirm that egocentric maps are highly effective when tasks require dynamic re-anchoring, spatial comparison, or structuring.

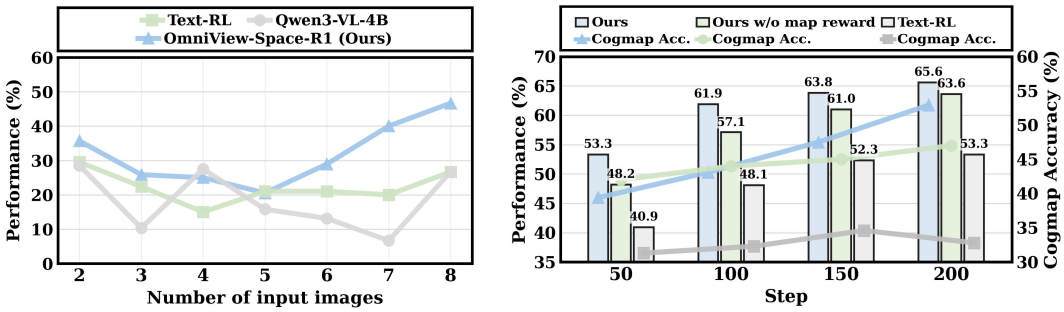

Figure 5: (a) Accuracy on MMSI-Bench by the number of input images. Tool-integrated RL remains more stable as views increase. (b) Training dynamics of answer accuracy and BEV quality. Image-cogmap supervision improves both metrics more consistently than text-cogmap supervision.

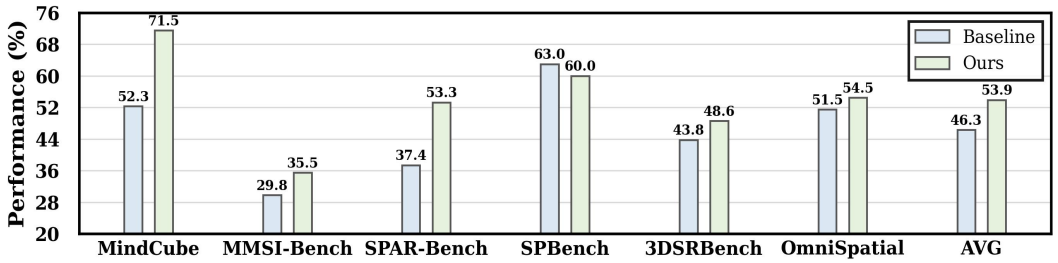

Figure 6: **Ablation across six benchmarks.** OmniView-Space broadly improves over the baseline.

4.3 Ablation Studies

Egocentric Cognitive Map vs. Text-Only Reasoning Baseline. Fig. 6 compares reasoning using multimodal egocentric cognitive maps against text-only approaches. OmniView-Space averages 53.9%, outperforming the text-only baseline of 46.3%. This 7.6-point improvement demonstrates that mapping observations into egocentric cognitive maps yields more reliable evidence than relying solely on the textual chain-of-thought. MindCube and SPAR-Bench show the largest gains (19.2 and 15.9 points, respectively), confirming the method's value in tasks requiring viewpoint re-anchoring and structured relational reasoning. The consistent improvements on MMSI-Bench, 3DSRBench, and OmniSpatial suggest this mapping strategy generalizes well to various multi-view settings. The exception is single-image SPBench, where the text baseline slightly outperforms OmniView-Space (63.0% versus 60.0%). This implies that existing language priors can adequately handle single-image tasks with limited viewpoint dependency, reducing the need for explicit spatial mapping.

Effectiveness of Multimodal Maps and Map Alignment Reward. As shown in Table 4, first, adding MPSM evidence as prompts improves over Qwen3-VL-4B, showing that query-aligned evidence is useful even without RL training. Second, map-alignment reward improves the distilled model from 46.4 to 48.7 on average, with gains on both MindCube and MMSI-Bench. This supports the role of ego-frame map supervision in producing more useful self-generated cognitive maps.

Table 4: **Ablation study.** We isolate the effects of MPSM evidence, tool-guided reasoning, evidence modality, and map-alignment reward.

Model	MindCube	MMSI	AVG.
Qwen3-VL-4B	35.9	27.7	31.8
+ MPSM-Map only	48.4	30.3	39.4
+ Multimodal MPSM	49.1	**31.7**	40.4
OmniView-Space-R1	**65.7**	31.6	**48.7**
- Map Alignment	63.6	29.1	46.4

Robustness to Number of Input Views. Fig. 5(a) reports the MMSI-Bench accuracy grouped by the number of input images. The two baselines, Text-RL and Qwen3-VL-4B, are competitive only for two-view inputs. Their accuracy fluctuates sharply and drops below 10% when given more than four images, revealing an inability to aggregate evidence across multiple viewpoints. Our tool-integrated RL method is substantially more stable across different image counts. It achieves 35.7% with two views, maintains 20% to 30% for three to six views, and increases to 40.0% and 46.7% for seven and eight views. Explicit cognitive-map tool calls appear to convert additional viewpoints into cumulative spatial evidence rather than distracting context, which explains the widening performance gap over the baselines as the input scale grows.

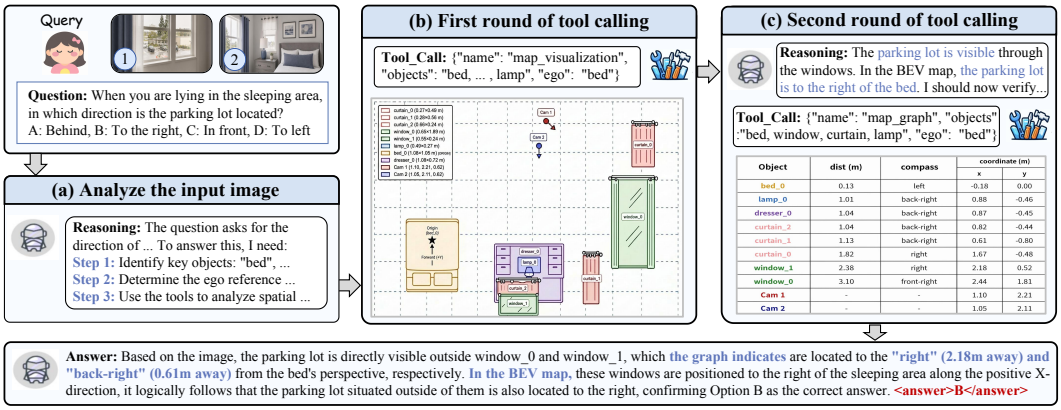

Figure 7: **Case study.** The model first identifies the query anchor and requests a bed-centered BEV map, which makes the window and parking-lot layout directly inspectable. It then calls the graph tool to verify ego-relative relations and grounds the final answer in the query-specified frame.

Cogmap Accuracy and Training Dynamics. Fig. 5(b) compares performance and Cogmap Acc. across training steps. Ours shows the greatest and most consistent improvement, with task performance increasing from 53.3% to 65.6% and Cogmap Acc. from 39.4% to 52.9%. Ours w/o map reward also improves in performance, from 48.2% to 63.6%, but its Cogmap Acc. grows more slowly and is surpassed by Ours in later training. Text-RL remains weaker on both axes: its performance rises only from 40.9% to 53.3%, while Cogmap Acc. stays around 31–35%. These trends indicate that the benefit does not come merely from introducing an intermediate representation. Rendering the egocentric cognitive map as an image gives the model a more faithful spatial state, and the map-alignment reward further links higher-quality self-generated maps with stronger reasoning.

Tool-Call Distribution Analysis. Fig. 8 analyzes how our model uses the cognitive map and graph tools on correctly answered samples. MindCube and MMSI-Bench show a balanced use of the two modalities, with image/graph ratios of 50/50 and 46/54, respectively. This matches their emphasis on qualitative viewpoint reasoning, where a visual map provides an intuitive layout for relative-position judgments. By contrast, 3DSRBench places more weight on graph evidence, with graph calls accounting for 60% of tool usage. This reflects the need for more explicit numerical or structural evidence in detailed 3D relation reasoning. Ultimately, this distribution suggests that different evidence forms are used for different demands.

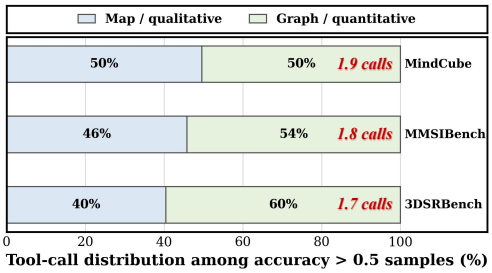

Figure 8: Tool-call distribution across benchmarks. The Visual Map is preferred for qualitative relative-position questions, whereas the textual graph excels in quantitative queries requiring precise numerical evidence.

Qualitative Analysis. Fig. 7 illustrates the OmniView-Space reasoning workflow. Based on the input views and spatial question, the model identifies relevant objects and the required reference frame to construct an egocentric cognitive map. This map explicitly represents the target frame, object layout, and relative directions within a visual BEV state. By using this representation as intermediate evidence, the model can accurately evaluate spatial relations from the correct viewpoint. This transparent process enhances interpretability, clarifying the rationale behind specific spatial judgments while mitigating the risk of relying on inconsistent reference frames.

5 Conclusion

We present OmniView-Space, a framework for multimodal egocentric reasoning that shifts spatial analysis from a default system perspective to a query-specified ego frame. To achieve this, we first propose the MPSM module constructs query-aligned visual BEV cognitive maps and textual spatial graphs. We further introduce a tool-guided policy that selects the required ego anchor and grounds answers based on MPSM. Finally, we distill this pipeline using MPSM trajectories and map rewards, enabling the model to reason through internally generated maps. Results confirm that query-aligned egocentric evidence effectively bridges metric geometry and language-based spatial reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, including improving clarity, grammar, and presentation, and for assisting with identifying and understanding recent related work to support a broader discussion of the literature. All AI-assisted outputs, including suggestions concerning related work, were carefully reviewed, verified against the original sources, and revised by the authors. The research ideas, methodology, experimental design, results, and conclusions were developed and validated by the authors. We take full responsibility for the final content of this work.

REFERENCES

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G Krishnan. Synthetic vision: Training vision-language models to understand physics. *arXiv e-prints*, pp. arXiv–2412, 2024.

Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. *arXiv preprint arXiv:2511.07403*, 2025.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025a.

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*, 2025b.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Fuhao Chang, Shuxin Li, Yabei Li, and Lei He. Vlm-3d: end-to-end vision-language models for open-world 3d perception. *arXiv preprint arXiv:2508.09061*, 2025.

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*, pp. 14455–14465, 2024a.

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. Spacetools: Tool-augmented spatial reasoning via double interactive rl. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37109–37120, 2026a.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. *arXiv preprint arXiv:2511.22659*, 2025a.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhenghao Chen, Huiqun Wang, and Di Huang. Egomind: Activating spatial cognition through linguistic reasoning in mllms. *arXiv preprint arXiv:2604.03318*, 2026b.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, 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.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025.

Yi Han, Enshen Zhou, Shanyu Rong, Jingkun An, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Lu Sheng, and Shanghang Zhang. Tiger: Tool-integrated geometric reasoning in vision-language models for robotics. *arXiv preprint arXiv:2510.07181*, 2025.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9590–9601. IEEE, 2024.

Yibin Huang, Wang Xu, Wanyue Zhang, Helu Zhi, Jingjing Huang, Yangbin Xu, Yangang Sun, Conghui Zhu, and Tiejun Zhao. Video2layout: Recall and reconstruct metric-grounded cognitive map for spatial reasoning. *arXiv preprint arXiv:2511.16160*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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. *arXiv preprint arXiv:2506.03135*, 2025.

Philip N Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.

Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9241–9251, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision*, pp. 95–111. Springer, 2026.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondences boost spatial-temporal reasoning in multimodal language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3792, 2025a.

Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025b.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025c.

Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia Sycara, and Yaqi Xie. pyspatial: Generating 3d visual programs for zero-shot spatial reasoning. *arXiv preprint arXiv:2603.00905*, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025a.

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. *arXiv preprint arXiv:2504.20024*, 2025b.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19446–19455, 2025.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. *arXiv preprint arXiv:2503.18470*, 2025.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. A new era of intelligence with gemini 3. *Google. URL: https://blog.google/products-and-platforms/products/gemini/gemini-3/ (accessed: 2026-01-16)*, 2025.

Rajarshi Roy, Devleena Das, Ankesh Banerjee, Arjya Bhattacharjee, Kousik Dasgupta, and Subarna Tripathi. Bydeway: Boost your multimodal llm with depth prompting in a training-free way. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6058–6064, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

Hengyi Wang, Ruiqiang Zhang, Chang Liu, Guanjie Wang, Zehua Ma, Han Fang, and Weiming Zhang. Allocentric perceiver: Disentangling allocentric reasoning from egocentric visual priors via frame instantiation. *arXiv preprint arXiv:2602.05789*, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9058–9069, 2025b.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. Visuothink: Empowering lvlm reasoning with multimodal tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21707–21719, 2025c.

Zehan Wang, Ziang Zhang, Jiayang Xu, Jialei Wang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything v2: Unifying orientation and rotation understanding. *arXiv preprint arXiv:2601.05573*, 2026b.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *Advances in Neural Information Processing Systems*, 38:143297–143330, 2026.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025a.

Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models' basic spatial abilities: A perspective from psychometrics. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11571–11590, 2025b.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

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*, 2025b.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 2026.

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*, 2025c.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data. *arXiv preprint arXiv:2506.03642*, 2025.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026a.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 70072–70141, 2026b.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Zhenfei Yin, Lijun Wang, and Huchuan Lu. Think3d: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026c.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. *arXiv preprint arXiv:2410.17385*, 2024.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025a.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025b.

Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix Overview

This appendix provides supplementary implementation details, audits, and analyses for OmniView-Space: Section B details the MPSM-guided SFT data construction and three-round filtering pipeline; Section C audits training–evaluation overlap across visual instances, question templates, and task families; Section D formulates the map-alignment rewards; Section E examines MPSM failure modes, fallback strategies, and adaptive tool routing; Section F analyzes the sensitivity of performance to rendered-map quality; Section G evaluates inference latency and computational cost-efficiency; Section H compares visual and textual cognitive maps via attention analysis; and the final sections discuss additional related work and limitations.

B MPSM-guided SFT Data Construction

The goal of the SFT data construction pipeline is to teach the model not only to answer spatial questions, but also to externalize an intermediate ego-frame cognitive map before answering. Each full cogmap trajectory is organized as:

$$\begin{aligned} &\texttt{<think>}\ R_1\ \texttt{</think>} \rightarrow \texttt{<cogmap>}\ J\ \texttt{</cogmap>} \\ &\rightarrow \texttt{<think>}\ R_2\ \texttt{</think>} \rightarrow \texttt{<answer>}\ \hat{a}\ \texttt{</answer>}. \end{aligned} \tag{7}$$

where J is a JSON-format ego-frame cogmap, R_1 explains how the cogmap is constructed from the images and the query, and R_2 performs forward reasoning from the cogmap to the final answer. We use Qwen3-VL-235B (Bai et al., 2025a) to produce the language rationales and MPSM to produce the intermediate visual/spatial evidence.

B.1 Initial Cogmap Trajectory Generation

For each raw sample with question q, image set $\mathcal{I}$, and ground-truth answer a^*, the pipeline first extracts task-relevant object hints and determines the query-specified ego reference. The ego reference can be camera-centric, object-centric, or direction-centric, following the same representation used by MPSM. The object hints and ego reference are then passed to the unified MPSM pipeline, which combines dense geometry, segmentation, and orientation estimation to generate a rendered BEV map and an AST-style spatial tree. The spatial tree is serialized into a JSON cogmap J containing the ego reference, object entries with image-space boxes and BEV footprints, and camera entries with BEV positions and facing angles.

After obtaining J, we generate two rationales. The first rationale R_1 explains the map construction process: which objects are relevant, why the selected ego frame is appropriate, and how the generated cogmap represents the scene. The second rationale R_2 is generated in a forward-reasoning mode: the model is given the question, images, and generated cogmap, and it derives a predicted answer $\hat{a}$ without being shown the ground-truth answer. The final supervised example is assembled in the format of Eq. 7. This construction is used for MindCube and SpatialLadder-derived samples, with dataset-specific preprocessing for extracting object hints from grounding annotations, spatial questions, or existing answers.

Hint and ego-anchor provenance. At test time and during tool-guided RL rollouts, the policy predicts both the object hints $\mathcal{H}$ and the ego anchor r from the question and input images. MPSM is not given ground-truth boxes, object labels, answers, or ego anchors in these settings. During SFT construction, available MindCube object annotations may be used to initialize object hints. For SpatialLadder- and VST-derived samples, object hints are obtained through Qwen3-VL-235B-assisted self-grounding. Ego-anchor annotations are unavailable, so the anchor is predicted for every SFT sample. These construction-time signals are used only to build training trajectories and are not exposed during evaluation or RL.

We further audit the supervision sources of the 26K CogMap-distillation SFT samples to quantify how often ground-truth annotations are available and how often self-grounding is required. Table 5 reports the breakdown. Ground-truth object annotations are available only for the MindCube-derived subset (22.1%), while the remaining 77.9% of samples rely on MLLM-assisted self-grounding. Since ground-truth ego-anchor annotations are unavailable for all sources, ego anchors are generated by Qwen3-VL-235B for every SFT sample. These statistics clarify that only a small fraction of the

Table 5: **Supervision sources of the CogMap-distillation SFT data.** Ground-truth object annotations are available only for the MindCube-derived subset, while ego anchors are self-generated for every sample because no ground-truth ego-anchor annotations exist.

SFT supervision path	Data source	# Samples	Percentage
GT object annotations	MindCube-derived	5,760	22.1%
MLLM-assisted self-grounding	SpatialLadder + VST-derived	20,273	77.9%
Total	All SFT samples	26,033	100%

training trajectories benefits from ground-truth object supervision, and none receives ground-truth ego anchors.

B.2 Three-round Filtering and Repair

The automatically generated trajectories can contain noise from segmentation failures, missing objects, imperfect ego-reference selection, or rationales that reach the right answer through shortcuts. We therefore use a three-round filtering and repair procedure for the full cogmap trajectories.

Round 1: forward-correct filtering. We first keep samples whose predicted answer $\hat{a}$ is already correct under the task metric. For multiple-choice questions, this requires exact option-letter agreement. For numerical questions, we use a thresholded relative-error score. These samples are retained directly because their rationales are generated in a forward manner from cogmap evidence to answer.

Round 2: answer-guided repair with critic filtering. For samples that fail Round 1, we ask whether the generated cogmap contains sufficient information to derive the ground-truth answer. If the cogmap is insufficient, the sample is deferred to Round 3. If the cogmap is sufficient, we generate a repaired R_2 conditioned on the ground-truth answer, but explicitly require the rationale to remain forward-style rather than saying that the answer is correct by assumption. A separate critic then checks for answer leakage, backward reasoning, circular reasoning, shortcut logic, and factual errors. Only critic-approved repaired trajectories are retained.

Round 3: missing-object refinement. For samples whose cogmap is insufficient, the most common failure mode is missing task-relevant objects. We identify objects mentioned by the answer options or question but absent from the cogmap, and use Qwen3-VL-235B to self-ground these missing objects with normalized 2D boxes. These supplemental detections are added to the original object hints, MPSM is rerun to produce an updated BEV map and spatial tree, and R_1, R_2, and $\hat{a}$ are regenerated. A sample is retained only if the regenerated answer is correct. This step recovers examples where the original MPSM reference was incomplete due to segmentation or object-filtering noise.

B.3 Quality Checks

After construction, we apply lightweight structural and accuracy checks before training. The format checker verifies the presence and nesting of the required tags, including `<think>`, `<cogmap>`, and `<answer>`. It also verifies that the cogmap JSON can be parsed and contains valid object and camera fields. For image-form trajectories, we check that input-image placeholders and rendered-BEV observations are consistent with the conversation structure. We also remove trajectories whose final answer fails the task metric, and flag abnormal samples with missing images, malformed boxes, repeated text, empty cogmaps, or excessive response length. These checks are used to reduce formatting noise before SFT.

B.4 Final SFT Mixture

The final SFT mixture combines direct-answer data, grounding-style cogmap data, and full cogmap-reasoning trajectories. Table 6 summarizes the sources. Overall, the mixture contains approximately 26K samples. The direct-answer subset maintains answer-level spatial reasoning ability, while the cogmap subsets teach the model to write structured ego-frame maps and to reason from those maps. The three-round filtering procedure is applied to the full cogmap-reasoning trajectories, while the

Table 6: **SFT data mixture.** We generate MPSM-guided cogmap trajectories from SpatialLadder, MindCube, and VST datasets.

Source	Size	Format / Filtering
SpatialLadder grounding	5K	Learn to predict the BEV Map directly from the input image through <think> R_1 </think> → <cogmap> J </cogmap> trajectories.
SpatialLadder spatial	7K	Multi-turn reasoning trajectories filtered from SpatialLadder (Li et al., 2025) after three-round filtering.
MindCube + VST	14K	Multi-turn reasoning trajectories filtered from MindCube (Yin et al., 2025) and VST-500K (Yang et al., 2025b) after three-round filtering.

Table 7: **Exact question-template overlap with training sources.** A zero count means no exact template match under this audit.

Training source → evaluation split	Samples	Exact matches
MindCube → MindCube-Tiny	1,050	333 (31.7%)
SpatialLadder → SPBench-SI	1,009	788 (78.1%)
SFT + RL → MMSI-Bench	1,000	0
SFT + RL → SPAR-Bench	7,211	0
SFT + RL → 3DSRBench	5,157	0
SFT + RL → OmniSpatial	1,533	0
SFT + RL → ViewSpatial-Bench	5,712	0

grounding-only SpatialLadder cogmap data is used directly because it supervises map construction rather than final answer derivation.

C Training–Evaluation Overlap Audit

We distinguish overlap in visual instances from overlap in question templates and task families. The cognitive-map distillation SFT mixture contains approximately 26K samples derived from MindCube, SpatialLadder, and VST; tool-guided RL uses approximately 21K samples from the MindCube and SpatialLadder training splits. Training MPSM trajectories are generated only from training samples: evaluation images, evaluation answers, and evaluation MPSM trajectories are not used for training.

Image and scene overlap. An audit of the SFT, RL, and evaluation indexes found no overlap in exact image paths, complete multi-view image bundles, or identifiable ScanNet scenes between training and evaluation. This disjointness does not imply that the question formulations or task families are disjoint.

Question-template overlap. Table 7 reports exact template matches between the training mixture and each evaluation split. For MindCube-Tiny and SPBench-SI, the comparison is against their corresponding training source; for the remaining benchmarks, it is against the SFT and RL mixtures. SPBench-SI denotes the 1,009-sample single-image subset of SPBench used in this audit.

Task-family overlap. MindCube-Tiny belongs to the MindCube family, while SPBench-SI is closely related to SpatialLadder in task format and construction. We therefore treat these two splits as in-domain evaluations despite their disjoint images. MMSI-Bench, SPAR-Bench, 3DSRBench, OmniSpatial, and ViewSpatial-Bench use different data sources and annotation pipelines and serve as external evaluations; their zero exact-template overlap does not rule out shared high-level spatial reasoning tasks.

As an additional check, Table 8 compares models trained directly on the public MindCube and SpatialLadder sources. Training on either source improves MindCube over the base model but does

Table 8: **Cross-benchmark performance of models trained on public spatial datasets.** Training on either source boosts MindCube accuracy over the base model but fails to transfer to the external MMSI-Bench.

Method	Training data	MindCube	MMSI-Bench
Base Model	None	37.6	28.6
MindCube-3B	MindCube-10K	51.7	23.9
SpatialLadder-3B	SpatialLadder-26K	43.4	27.4

not improve MMSI-Bench, suggesting that exposure to those training sources alone need not transfer to this external benchmark.

By contrast, the main evaluation reports an MMSI-Bench increase from 27.7 for Qwen3-VL-4B to 35.5 for OmniView-Space (Table 1), alongside improvements on SPAR-Bench, 3DSRBench, and OmniSpatial. The broader results are consistent with transfer beyond the two in-domain task families, although the training-source comparison alone cannot isolate the effect of query-aligned representations from all other training differences.

D Map Alignment Reward Details

This section expands the map-alignment reward used in the cognitive-map distillation stage. The model predicts a JSON-format cogmap $\hat{J}$ that contains an ego reference, a set of object entries, and a set of camera entries. Each object entry contains a predicted BEV footprint $\hat{b} = [\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2]$ and its BEV center $\hat{p} = ((\hat{x}_1 + \hat{x}_2)/2, (\hat{y}_1 + \hat{y}_2)/2)$. Each camera entry contains a BEV position and a facing angle. We compare this prediction against MPSM-generated scene-cache references expressed in the same ego frame.

D.1 Soft Footprint-Size Consistency

The object reward uses a soft size-consistency multiplier in addition to center distance. Given a predicted BEV box $\hat{b}$ and a reference BEV box b^*, let $(\hat{w}, \hat{h})$ and (w^*, h^*) be their widths and heights. The size score is

$$s(\hat{b}, b^*) := f + (1 - f) \exp\left(- \frac{|\log(\hat{w}/w^*)| + |\log(\hat{h}/h^*)|}{2} \right), \tag{8}$$

where $f = 0.5$ in our implementation. Thus $s \in [f, 1]$: center distance remains the dominant signal, while severe footprint-size mismatch is softly penalized. If either box is missing or degenerate, we set $s = 1$ so that incomplete size information does not invalidate the center-distance reward.

D.2 Filtered Object Alignment Reward R_{map}

R_{map} compares predicted objects with the filtered MPSM reference, which contains question-relevant objects selected from the scene cache. All coordinates are expressed in the query-specified ego BEV frame. If the ego reference is itself an object, that ego object is removed from both prediction and reference before scoring, because it is always placed at the ego origin and would otherwise create a trivial zero-distance match.

Objects are matched category by category. Let $\mathcal{C}$ be the set of object categories after removing instance suffixes, e.g., `chair_2` and `chair_3` share the category `chair`. For category c, let $\mathcal{P}_c$ be the predicted objects and $\mathcal{G}_c$ be the reference objects. We form a cost matrix using BEV center distances and use Hungarian matching to obtain a one-to-one matching $\mathcal{M}_c \subseteq \mathcal{P}_c \times \mathcal{G}_c$. Let $n_c = |\mathcal{G}_c|$, $\hat{n}_c = |\mathcal{P}_c|$, and $N_o = \sum_{c \in \mathcal{C}} n_c$. The reference scene diagonal is

$$D_o = \sqrt{(x^*_{\max} - x^*_{\min})^2 + (y^*_{\max} - y^*_{\min})^2}, \tag{9}$$

computed from all reference BEV centers, with a fallback lower bound to avoid a degenerate normalizer. For a matched pair $(i, j) \in \mathcal{M}_c$, where $\hat{p}_i$ and p^*_j are the predicted and reference BEV

centers, the pair score is

$$r_{ij}^{\text{obj}} := \left[1 - \frac{\|\hat{p}_i - p_j^*\|_2}{D_o}\right]_+ s(\hat{b}_i, b_j^*), \tag{10}$$

where $[x]_+ = \max(x, 0)$. The category score averages over the larger of the predicted and reference counts, so missing or extra objects reduce the score:

$$R_c = \frac{1}{\max(n_c, \hat{n}_c)} \sum_{(i,j)\in\mathcal{M}_c} r_{ij}^{\text{obj}}. \tag{11}$$

Finally, categories are weighted by their reference frequency:

$$R_{\text{map}} = \sum_{c\in\mathcal{C}} \frac{n_c}{N_o} R_c. \tag{12}$$

This reward therefore encourages the model to predict the correct object categories, place them at the correct ego-frame BEV locations, and preserve approximate footprint sizes.

D.3 Camera Alignment Reward R_{cam}

R_{cam} supervises the camera entries in the predicted cogmap. It evaluates both camera positions and facing directions in the query-specified ego BEV frame. If the ego reference is a camera, that ego camera is excluded from per-camera position/orientation scoring because its location is fixed at the origin by construction; however, it can still participate in pairwise direction comparisons. If the ego reference is an object or direction, no camera is excluded.

Let $\hat{c}_i$ and c_i^* be the predicted and reference BEV positions of camera i, and let $\hat{\theta}_i$ and θ_i^* be their facing angles. The camera-position normalizer is computed from reference camera positions only,

$$D_c = \max\left(\sqrt{(x_{\max}^* - x_{\min}^*)^2 + (y_{\max}^* - y_{\min}^*)^2}, D_{\min}\right), \tag{13}$$

where $D_{\min} = 0.3$ meters. The lower bound avoids unstable rewards in pivot-like scenes where cameras are nearly co-located.

The per-camera distance score is

$$r_{\text{dist}}^i = \left[1 - \frac{\|\hat{c}_i - c_i^*\|_2}{D_c}\right]_+. \tag{14}$$

To capture left–right and front–back layout, we also compute a signed pairwise direction score. For all valid camera pairs (i, j), let $\Delta\hat{c}_{ij} = \hat{c}_j - \hat{c}_i$ and $\Delta c_{ij}^* = c_j^* - c_i^*$. The pairwise term is

$$r_{\text{pair}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{\Delta\hat{c}_{ij}^\top \Delta c_{ij}^*}{\|\Delta\hat{c}_{ij}\|_2 \|\Delta c_{ij}^*\|_2}, \tag{15}$$

where $\mathcal{P}$ is the set of non-degenerate camera pairs. This score is signed in $[-1, 1]$, so a left–right flip can actively reduce the reward rather than merely receiving a smaller positive score. The final position reward for camera i is

$$r_{\text{pos}}^i = \text{clip}\left(\rho r_{\text{pair}} + (1-\rho) r_{\text{dist}}^i, -1, 1\right), \tag{16}$$

with $\rho = 0.7$ in our implementation. The orientation reward is cosine angular similarity,

$$r_{\text{ori}}^i = \frac{1 + \cos(\hat{\theta}_i - \theta_i^*)}{2}, \tag{17}$$

where the angle difference is wrapped to $[-180^\circ, 180^\circ)$. A 90° error gives 0.5, and a 180° error gives 0.

The final camera reward is

$$R_{\text{cam}} = \frac{1}{N_c} \sum_{i\in\mathcal{K}} \left(\alpha_{\text{pos}} r_{\text{pos}}^i + \alpha_{\text{ori}} r_{\text{ori}}^i\right), \tag{18}$$

where $\mathcal{K}$ is the set of matched cameras, N_c is the number of reference cameras to be scored, and missing cameras contribute zero through the denominator. We use $\alpha_{\text{pos}} = 0.6$ and $\alpha_{\text{ori}} = 0.4$.

Table 9: **MPSM toolkit failure audit.** Rates are computed on MMSI-Bench and MindCube-Tiny.

Failure proxy	MMSI-Bench	MindCube-Tiny
Overall MPSM toolkit failure	6.2%	4.1%
Depth / reconstruction failure	0.9%	0.6%
Segmentation failure	3.2%	2.1%
Object-orientation failure	1.2%	0.9%

E MPSM Failure Modes, Fallbacks, and Applicability

We audit the MPSM pipeline on the full evaluated splits of MMSI-Bench (1,000 samples) and MindCube-Tiny (1,050 samples). An overall toolkit failure means that MPSM cannot produce a usable CogMap from the policy-predicted object hints and ego anchor because of external geometry errors, including reconstruction, segmentation, or BEV projection. A segmentation failure means that none of the visible target objects is segmented or restored to a valid object area; an orientation failure means that an object-centric query lacks a valid object-facing estimate. Table 9 reports the audited failure proxies. The stage-wise proxies need not exhaust or sum to the overall failure rate.

Handling low-confidence or incomplete maps. Low-quality MPSM outputs arise mainly from three sources: low-confidence reconstruction, failed object segmentation, and unreliable orientation estimation. We address each as follows.

Low-confidence reconstruction. MPSM applies confidence filtering before constructing the CogMap. Only pixels and frames that pass the pixel-level and frame-level confidence gates in Eq. 2 are used for 3D reconstruction and BEV projection, which prevents unreliable geometry from corrupting the map; the audited scene-level reconstruction success rate is close to 99%. Moreover, even when some object-level modules subsequently fail, the CogMap can still preserve the camera-pose distribution, which has been shown to benefit spatial understanding (Yang et al., 2026).

Failed object segmentation. If task-relevant objects mentioned in the question or candidate answers are missing from the CogMap during SFT trajectory construction, Qwen3-VL-235B localizes them with 2D boxes and MPSM is rerun with the additional object hints, as described in the three-round repair procedure above (Sec. B). During RL, if a policy-predicted object is absent from the reference map because of segmentation or object-filtering errors, its policy-predicted 2D box is used only as a fallback: cached 3D points are projected into the corresponding image, points within the box are selected, and their BEV extent is used as the object's footprint.

Unreliable orientation estimation. Orientation-estimation failures are rare (1.2%; Table 9). When they occur for an object-centric query, MPSM falls back to a camera-centric or global CogMap rather than imposing an unreliable object-centric frame, which keeps the spatial evidence usable. These fallbacks preserve partial spatial evidence but do not guarantee that the resulting map is correct.

Task regimes where MPSM is less beneficial. MPSM is designed for tasks requiring viewpoint transformation, cross-view aggregation, or ego-frame re-anchoring, and is less helpful for questions answerable directly from 2D perception. Specifically, (1) on metric-heavy queries in single-image or sparse-view settings (e.g., absolute distance or object size estimation), monocular 3D reconstruction is inherently under-constrained and BEV projection can introduce metric noise; and (2) on fine-grained appearance queries, the compressed BEV sketch omits pixel-level visual details. This explains why our model slightly trails the text-only baseline on the single-image SPBench-SI split: over 60% of SPBench-SI questions are metric-related under a single view, whereas our RL policy is primarily trained on multi-view scenarios requiring cross-view evidence aggregation.

Avoiding unnecessary map construction. Our prompt specifies suitable scenarios for each tool, and outcome-based RL encourages the policy to invoke MPSM only when spatial evidence improves final correctness. To quantify this routing behavior, we audit the Attribute subset of MMSI-Bench (130 samples), which mixes appearance-oriented and spatially grounded questions. Each sample is annotated as either *MPSM-beneficial* or *direct-reasoning sufficient*. As shown in Table 10, the policy matches the expected routing in 79.2% of cases and skips MPSM in 75.0% of direct-reasoning cases, improving subset accuracy from 26.0% to 28.7%.

Table 10: **Tool-routing audit on the MMSI-Bench Attribute subset.** The policy selectively invokes MPSM when spatial evidence is beneficial and largely skips it when direct visual perception suffices.

Sample type	# Samples	Expected behavior	Policy agreement
MPSM-beneficial	42	Call MPSM	88.1%
Direct-reasoning sufficient	88	Skip MPSM	75.0%
Overall	130	Match human routing label	79.2%

Table 11: **Sensitivity to rendered-map quality on MindCube.** The accuracy gain of our method is larger for high-quality CogMaps, yet remains positive even for low-quality CogMaps.

Map quality	# Samples	Text-RL	Ours	Δ
All	1,050	53.24	65.71	+12.48
High-quality CogMap	525	54.05	72.19	+18.14
Low-quality CogMap	525	52.43	59.24	+6.81

Table 12: **Inference latency on MMSI-Bench.** Latency cost is measured per 1% accuracy gain over random guess (25.0%).

Method	Think / Round	Tool / Round	Total Latency	Acc.	Latency Cost Per 1% Gain ↓
Random-Guess	–	–	–	25.0	–
Qwen3-VL-235B-Think (end-to-end)	–	–	31.76s	32.6	4.18s
Text-only CoT RL	–	–	7.58s	29.8	1.58s
Ours with distillation	4.12s	0.29s	8.53s	31.6	1.29s
Ours	5.46s	1.64s	11.98s	35.5	1.14s

Table 13: **Per-round tool-call latency breakdown of the integrated version.**

Metric	DA3	SAM3	OriAny	All Tools
Time / Round	0.50s	1.05s	0.08s	1.64s

F Sensitivity to Rendered-Map Quality

We examine whether performance depends on the quality of the generated map. We quantify rendered-map quality on MindCube by comparing each generated CogMap against its ground-truth CogMap annotation. We then rank samples by CogMap accuracy and split them into two equal halves: the top 50% as high-quality CogMaps and the bottom 50% as low-quality CogMaps. Table 11 reports the accuracy of Text-RL and our model on each group.

The accuracy gain is substantially larger for high-quality CogMaps (+18.14) than for the low-quality half (+6.81), indicating that better rendered maps contribute to stronger spatial reasoning. Crucially, even the low-quality CogMap group outperforms Text-RL by +6.81 points, demonstrating that imperfect maps remain beneficial by supplying valuable 2D spatial priors, such as coarse scene layouts and camera-pose distributions.

This analysis is correlational rather than causal, and one confound is that easier scenes may both yield higher-quality maps and be easier to answer correctly, which would inflate the gap between the high- and low-quality groups. Because samples are grouped by the quality of the model's own predicted map rather than by an exogenous intervention, we cannot fully separate this scene-difficulty confound from the effect of map quality. We note, however, that the comparison against Text-RL is within each group: even on the low-quality half, where per-sample map quality is poor, our model still exceeds Text-RL by +6.81 points, and Text-RL has access to the same questions and images. This within-group comparison is less exposed to the scene-difficulty confound than the high-versus-low contrast, since both methods are evaluated on the identical sample set. A controlled study that fixes scene difficulty while varying only map quality would be required to establish causality, and we leave this to future work.

Table 14: **CogMap format and visual attention.** Attention is averaged from generated reasoning tokens to image tokens on sampled MindCube cases.

Method	CogMap form	Image attention	MindCube acc.
Text-RL	Text CogMap	0.0037	53.2%
OmniView-Space	Image CogMap	0.0425 ($\times$11.5)	65.7%

G Inference Latency and Computation Costs

We measure inference latency on MMSI-Bench using NVIDIA H20 GPUs and report both total latency and the latency cost per 1% accuracy gain over random guess (25.0%). Table 12 summarizes the results. Although the tool-integrated version introduces external MPSM calls and increases total latency compared with text-only CoT RL (11.98s vs. 7.58s), it provides a larger accuracy gain (35.5% vs. 29.8%). To quantify this tradeoff, we report the latency cost per 1% accuracy gain over random guess. OmniView-Space achieves the best cost-efficiency (1.14s per 1% gain), compared with 1.58s for text-only CoT RL and 4.18s for Qwen3-VL-235B-Think. The distilled variant further reduces total latency to 8.53s while retaining part of the tool-guided gain, providing a cheaper deployment option. We also break down the tool-call latency in the integrated version in Table 13. Thus, the integrated model provides stronger spatial grounding and accuracy, while the distilled model offers a cheaper inference option with competitive performance.

H Visual versus Textual Cognitive Maps

The visual and textual CogMaps encode spatial evidence through different interfaces. We hypothesize that image-CogMap supervision provides stronger visual grounding. A text-CogMap serializes coordinates and relations into tokens interpreted by the autoregressive language model, which may increase reliance on language priors and permit ungrounded relation guesses in complex layouts (Yang et al., 2026). In contrast, an image-CogMap retains the BEV layout as visual evidence alongside the original views. This may encourage the model to cross-check the map against the input images, reducing language-prior reliance and spatial hallucinations. The visual layout also preserves global topology: relative positions and connectivity can be inspected without reconstructing them from serialized coordinates. These are hypotheses about the observed performance gap, not effects isolated by the attention measurement below.

Attention analysis. We measured the average attention from generated reasoning tokens to image tokens on sampled MindCube cases under the text-CogMap and image-CogMap conditions. Table 14 reports the attention values together with the reported MindCube accuracies for these conditions. The attention score increases from 0.0037 to 0.0425 (approximately 11.5-fold using the rounded values), alongside an increase in accuracy from 53.2% to 65.7%. The attention sample should not be confused with the evaluation split used for the accuracy figures.

Higher image-token attention is consistent with stronger use of visual evidence, but does not by itself show that the map format caused the accuracy gain, that the model explicitly cross-checked the original images, or that hallucinations decreased. The topology hypothesis is also consistent with the reported VSI-Bench improvements over Qwen3-VL-4B on Relative Direction (+10.2 points) and Route Planning (+14.2 points) when seven frames are uniformly sampled per video. These results involve the overall method rather than a controlled image-versus-text CogMap comparison, so they do not isolate the map format. The textual graph remains useful when a question requires explicit distances, attributes, or support relations.

I More discussion with related work

Several recent methods also combine external tools with vision-language models, but they differ from OmniView-Space in both the target setting and the supervision applied to intermediate evidence. Visual Program Distillation (VPD) (Hu et al., 2024) constructs verified visual-program trajectories with external tools and distills them into a vision-language model through supervised training. Its primary objective is to imitate tool-produced programs and their reasoning traces. SpaceTools (Chen et al., 2026a) uses double interactive reinforcement learning to invoke spatial

tools and provide additional cues for spatial reasoning, primarily for local or single-image inputs. In contrast, OmniView-Space targets multi-image, viewpoint-dependent reasoning: MPSM fuses evidence across views, selects a camera-, object-, or direction-centric ego anchor from the query, and returns a visual BEV map and a textual spatial graph expressed in that query-required frame. The distinction also concerns what is optimized after tool use. OmniView-Space uses MPSM to construct SFT trajectories, but subsequently applies map-alignment rewards during GRPO to evaluate the model's own generated cognitive map, including object locations, object footprints, camera positions, and camera orientations in the selected ego frame. Therefore, the model is trained not only to reproduce a successful tool trajectory, but also to construct spatial evidence whose geometry is aligned with the reference frame required by the question. The contribution is thus the query-aligned intermediate representation and its geometric supervision for multi-view reasoning, rather than the use of external tools, reinforcement learning, or distillation in isolation.

J LIMITATIONS

One potential limitation is that the tool-augmented pipeline still depends on the quality of depth estimation, segmentation, and object-orientation estimation. Visual truncation, occlusion, or missing task-relevant objects can lead to incomplete egocentric evidence and incorrect spatial judgments. Although the internalized model reduces reliance on external geometry tools at inference time, its self-generated cognitive maps can still drift geometrically, especially for out-of-domain layouts, sparse views, or ambiguous object orientations. These limitations suggest several directions for future work, including more robust open-vocabulary spatial mapping, stronger map-level supervision, and reasoning policies that can detect when their self-generated maps are unreliable. More broadly, we view ego-frame grounding as a step toward MLLMs that not only describe visual scenes, but also re-instantiate them from the viewpoint required by a task.