

# Dense Reward for Multi-View 3D Reasoning with Global Maps and Local Views

Jiho Choi<sup>1\*</sup>, Seonho Lee<sup>2\*</sup>, Seojeong Park<sup>1</sup>, and Hyunjung Shim<sup>1</sup>

<sup>1</sup> Graduate School of Artificial Intelligence, KAIST, Republic of Korea

<sup>2</sup> KRAFTON, Republic of Korea

{jihochoi, seojeong.park, kateshim}@kaist.ac.kr, {glanceyes}@krafton.com

**Abstract.** Multi-view 3D Visual Question Answering (MV3D-VQA) requires integrating partial observations into a coherent 3D scene representation and selecting informative viewpoints for multi-step spatial reasoning. However, current multimodal LLMs are typically trained with sparse, answer-level supervision, which often yields inconsistent cross-view reasoning and brittle view selection. We present **DR-MV3D** (Dense Reward for MV3D-VQA), a map-grounded learning framework that provides dense, verifiable rewards to supervise the reasoning process. Our approach decomposes MV3D-VQA into (i) allocentric global map construction, (ii) question-conditioned view-trajectory planning, and (iii) egocentric grounding for answer prediction. To make intermediate steps learnable without manual annotations, we introduce two rewards: a global consistency reward that aligns the predicted map with geometry-consistent pseudo targets from frozen 3D vision foundation models (e.g., VGGT + SAM3), and a local trajectory reward that supervises ordered viewpoint selection. We optimize the full pipeline with trajectory-level policy optimization (GRPO). Experiments on MindCube, VSI-Bench, and BLINK (MV) show that DR-MV3D consistently improves over strong multi-image baselines, supporting the effectiveness of process-level dense supervision for multi-view 3D reasoning.

Code is available at: <https://github.com/kaist-cvml/dr-mv3d>

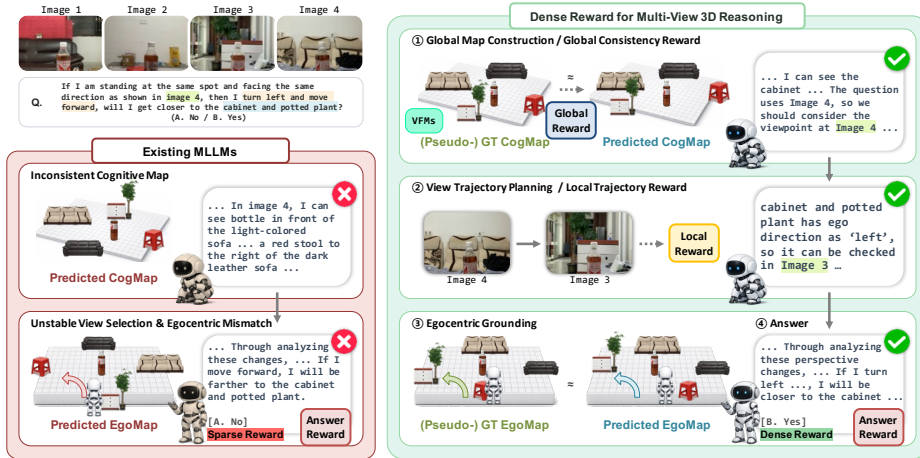
**Keywords:** Multi-view VQA · Reasoning MLLM · Dense Rewards

## 1 Introduction

Recent advances in multimodal large language models (MLLMs) [20,30,36,46,51] have achieved remarkable progress on visual perception, visual question answering, and multimodal reasoning [6,14,46]. However, many real-world environments are inherently three-dimensional and can only be observed through multiple partial viewpoints. This has motivated growing interest in *multi-view 3D visual question answering* (MV3D-VQA) [17,59], where a model must integrate observations from multiple viewpoints and reason about spatial relationships within

---

\* Equal contribution.



**Fig. 1: Overview of DR-MV3D.** Existing MLLMs trained with *sparse*, answer-level supervision (left) often build inconsistent cognitive maps and misread egocentric directions, leading to wrong answers despite plausible-looking reasoning. In contrast, DR-MV3D (right) supervises the reasoning process with *dense*, verifiable rewards: a *global reward* aligns the predicted allocentric cognitive map with a geometry-consistent target from a frozen 3D vision foundation model (VFM), and a *local trajectory reward* guides ordered viewpoint selection so that the model grounds the query in the correct view (e.g., resolving the ego-direction “left” via the appropriate image) and arrives at the correct answer.

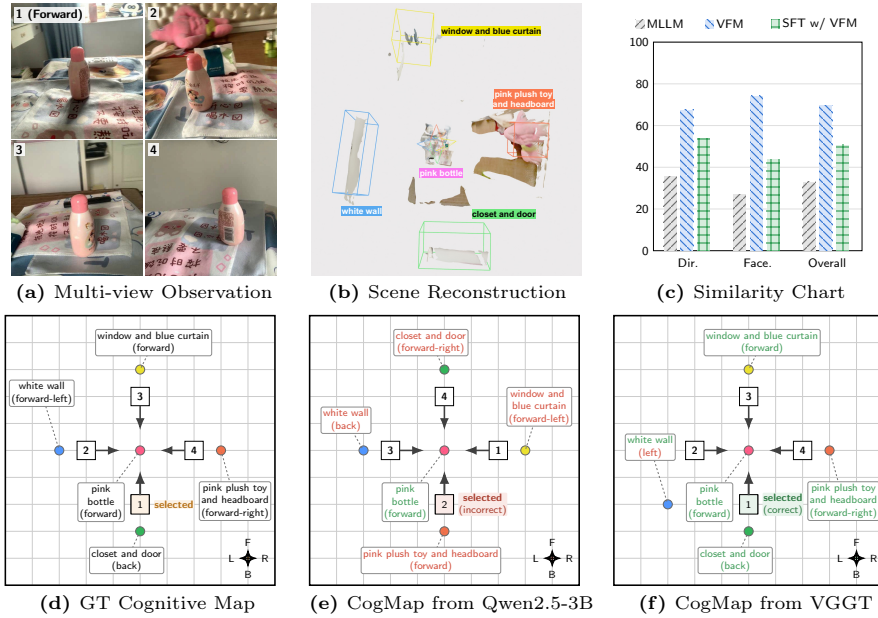
a scene. Unlike conventional visual question answering that relies on a single image, MV3D-VQA requires aggregating fragmented visual evidence across views to form a coherent understanding of the underlying 3D scene [12, 17, 44, 59]. Such spatial reasoning is essential for embodied and physical intelligence [10, 22, 62], where agents must operate under partial observability and reason about object relations, occlusions, and hypothetical movements.

Despite these advances, existing MLLMs still struggle to construct and utilize coherent *three-dimensional spatial representations* from partially observable multi-view inputs [12, 18, 44, 52, 59]. Empirical studies on multi-view reasoning benchmarks reveal several systematic failure modes: models frequently produce inconsistent predictions across viewpoints, exhibit unstable reasoning under occlusion, and perform close to random on compositional spatial queries that require integrating evidence from multiple views [12, 17, 44, 59]. These observations suggest that the core challenge is not merely the availability of visual inputs, but the difficulty of constructing and maintaining a consistent scene-level representation that supports structured spatial reasoning. Although MLLMs may implicitly encode partial spatial knowledge, they often fail to align, retrieve, and update this information coherently across viewpoints, preventing fragmented observations from being organized into a unified representation of the underlying 3D scene.

Recent work has begun to address this limitation by introducing explicit intermediate representations. MindCube [44] and related studies [15,17,59] demonstrate that providing additional views alone yields limited gains; instead, they show that performance improves when models first construct an intermediate representation such as a *cognitive map* and then reason over it. These findings establish the *map-then-reason* paradigm as a promising direction: the cognitive map functions as a persistent spatial workspace that organizes fragmented observations into a coherent scene-level structure for downstream tasks such as global scene summarization (spatial memory) and question-conditioned view planning.

Despite this progress, several key challenges remain. First, most existing approaches rely solely on allocentric (world-centered) representations. While allocentric maps provide a stable global reference frame, reasoning tasks often involve egocentric queries such as viewpoint-dependent relations or hypothetical movements, creating a representation misalignment between the constructed map and the reasoning process. Second, MLLMs often struggle to produce geometrically consistent 3D structures even when generating intermediate maps. In contrast, specialized 3D Vision Foundation Models (VFMs) [43,45], such as VGGT, exhibit strong capabilities in multi-view geometric reasoning, including depth estimation, pose consistency, and object-level spatial layout reconstruction. This reveals a clear representation quality gap between language-driven map generation and geometry-grounded vision models. As illustrated in Figure 2, cognitive maps derived from VGGT [43] are more similar to the ground-truth cognitive map than those generated by Qwen2.5-3B [36], suggesting that VFM-based representations can provide more reliable structural signals for spatial reasoning. Third, most existing reinforcement learning approaches rely primarily on sparse rewards defined by final answer correctness, which provides limited guidance for learning spatially structured reasoning processes.

In this work, we propose DR-MV3D (Dense Reward for Multi-View 3D reasoning), a map-grounded multi-view reasoning framework that explicitly addresses these challenges as in Figure 1. To resolve the allocentric-egocentric representation misalignment (Challenge 1), our approach constructs a global *allocentric cognitive map* summarizing scene-level spatial structure from multi-view observations, and then grounds the acquired evidence into viewpoint-aligned *egocentric representations* that match the reasoning dynamics of MLLMs. To bridge the geometric reliability gap in language-driven map generation (Challenge 2), we incorporate a 3D Vision Foundation Model (VFM) as a structural prior to regularize cognitive map construction toward geometrically consistent 3D structure. Conditioned on the global map and the input question, the model further plans a question-aware *view trajectory* that selectively acquires evidence from informative viewpoints. Finally, to alleviate sparse credit assignment in long-horizon multi-view reasoning (Challenge 3), DR-MV3D optimizes the entire pipeline through trajectory-level policy optimization with a multi-level reward that is *dense along the trajectory* (i.e., step-wise feedback for each view/action) [4,58], jointly supervising map consistency, informative view selection, and final answer correctness. To further improve the geometric reliability of the learned spatial



**Fig. 2: Cognitive map and metric comparison.** Top: multi-view observations, VFM-based scene reconstruction (VGGT+SAM3), and similarity scores to the GT cognitive map. The similarity is reported from three aspects: *Dir.* measures inter-object directional relations (relative spatial ordering between objects), *Face.* measures the facing orientation of the viewpoint, and *Overall* is a weighted combination of *Dir.* and *Face.* We compute *Dir.* as pairwise directional agreement and *Face.* as viewpoint-facing agreement; *Overall* is their weighted average. Bottom: GT cognitive map, an MLLM-generated cognitive map (Qwen2.5-3B), and a VFM-derived cognitive map. The VFM-derived cognitive map achieves higher similarity to GT than the MLLM-generated map, suggesting that VGGT+SAM3 can provide a reliable pseudo-structural signal when GT cognitive maps are unavailable.

representation, we additionally incorporate supervision from a pretrained 3D Vision Foundation Model. Specifically, we utilize geometry-consistent scene representations produced by models such as VGGT [43] combined with SAM3 [3] as pseudo-structural targets. These signals softly align the MLLM-generated cognitive maps with physically plausible 3D layouts without requiring manual annotations.

Our contributions are summarized as follows. We propose **DR-MV3D**, a map-grounded multi-view reasoning framework that integrates observations into global allocentric representations and subsequently aligns them with egocentric reasoning dynamics. To address the limitations of sparse supervision, we introduce a dense reward learning scheme that utilizes pretrained vision foundation models to provide step-wise feedback for both scene consistency and informative viewpoint selection. Extensive evaluations demonstrate the effectiveness of

DR-MV3D, where our 3B model achieves competitive performance across multiple benchmarks, notably reaching 66.5% accuracy on MindCube-Tiny, a 28.7%p improvement over the baseline (Qwen2.5-3B [36]), while also delivering superior results on VSI-Bench (37.1 Avg) and BLINK (MV) (56.4). These findings highlight that dense structural supervision provides a more effective learning signal for multi-step spatial reasoning than standard answer-level optimization.

## 2 Related Work

**3D Understanding and Spatial Reasoning in VLMs/MLLMs.** Recent benchmarks and analyses show that, despite strong 2D perception and language-driven reasoning, MLLMs remain brittle in 3D understanding under partial observability and multi-view evidence [12, 15, 17, 28, 44, 52]. Multi-image or multi-view evaluations highlight recurring failures in viewpoint consistency, relative spatial relations, and spatial memory, suggesting that correct answers can stem from shortcuts rather than coherent scene-level representations [12, 15, 44, 52, 53].

Prior work addresses these limitations in two complementary ways: (i) injecting 3D-aware representations, such as learning compact 3D scene abstractions from multi-view images, aligning multimodal features with pretrained 3D foundations, or distilling geometric knowledge from 3D vision foundation models into VLMs [17, 19, 24, 25, 59], and (ii) improving 3D reasoning behavior via explicit reasoning traces/CoT supervision, perspective-aware mental imagery, or multi-reward RL for long-horizon spatial reasoning and view selection [18, 24, 29, 31, 56]. However, such supervision is often sparse (answer-level), relies on task-specific intermediate annotations, or distills geometric priors only into static feature representations rather than supervising the multi-step reasoning process itself. Our work complements these efforts by introducing *dense rewards* that directly shape multi-view spatial reasoning: a global reward for cross-view scene consistency and a local reward for informative viewpoint selection, enabled by scalable pseudo-structural supervision from pretrained vision foundation models.

**Verifiable Rewards and GRPO for Multimodal Reasoning.** A growing body of work replaces human preference supervision with verifiable rewards [14, 38, 41, 49], where learning signals are computed by objective checkers (e.g., exact-match correctness, rule-based constraints, or metric-driven evaluators). In this paradigm, GRPO [38] has emerged as a practical reinforcement learning (RL) optimization strategy that avoids an explicit value/critic model by estimating baselines from group-wise samples, enabling scalable RL for reasoning behaviors. Building on this idea, DeepSeek-R1 [14] demonstrates that strong long-horizon reasoning can be induced primarily from correctness-based, automatically verifiable rewards, reducing dependence on curated chain-of-thought supervision.

Recent extensions bring verifiable-reward RL to vision-language models by designing rewards that evaluate intermediate reasoning steps and grounded perception outputs. R1-VL [58] proposes step-wise GRPO with rule-based dense rewards for multi-step multimodal reasoning, while Vision-R1 [57] and Visual-RFT [34] explore vision-guided or metric-driven reinforcement fine-tuning to

align visual reasoning and perception without human feedback. Complementary directions improve generalization via curriculum RL [8], two-stage rule-based RL for compact MLLMs [35], iterative self-improvement [9], and cross-modal formalization for robust multimodal reasoning [54]. Related efforts also leverage reasoning-aware training to produce grounded outputs in segmentation and referring tasks [23, 32, 33]. Despite these advances, most prior verifiable-reward frameworks predominantly target single-view 2D settings and do not explicitly enforce cross-view spatial consistency under partial observability, which is central to multi-view 3D VQA.

### 3 Proposed Method

#### 3.1 Problem Definition

We consider multi-view 3D visual question answering (MV3D-VQA) [12, 17, 44], where a scene is observed through a finite set of viewpoints. Formally, given images  $\mathcal{I} = \{I_1, \dots, I_N\}$  captured from different viewpoints of the same scene and a spatial reasoning question  $q$ , the goal is to predict the correct answer  $a^*$ .

Unlike conventional 2D VQA [1, 13], MV3D-VQA requires the model to actively aggregate fragmented information across space. We formulate the problem as trajectory-level policy optimization to reflect this active reasoning process [48, 55], where the model must sequentially determine which observations are necessary to resolve incomplete spatial evidence. The MLLM induces a structured trajectory  $\tau$  that first builds a coarse allocentric cognitive map  $\mathcal{C}^{\text{allo}}$ , then selects a view trajectory  $\mathcal{V} = (v_1, \dots, v_T)$  conditioned on  $(\mathcal{C}^{\text{allo}}, q)$ . For each selected view  $v_t$ , it constructs a corresponding egocentric map  $\mathcal{C}_t^{\text{ego}}$ , yielding an egocentric map sequence  $\mathcal{C}_{1:T}^{\text{ego}}$  that supports continued reasoning and the final answer  $\hat{a}$ . The task objective is to maximize expected answer-level reward:

$$\max_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}(\cdot | \mathcal{I}, q)} [R_{\text{ans}}(\hat{a}(\tau), a^*)], \quad R_{\text{ans}} = \mathbb{I}[\hat{a}(\tau) = a^*], \quad (1)$$

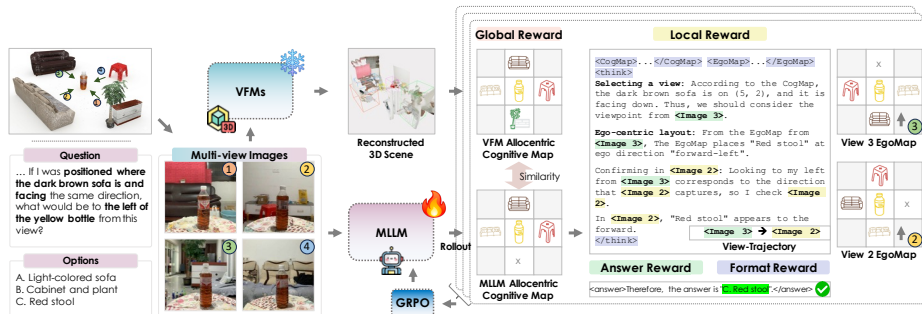
where  $\hat{a}(\tau)$  is the predicted answer from trajectory  $\tau$ , and  $\mathbb{I}[\cdot]$  is the indicator function.

#### 3.2 Overall Framework

Our framework, **DR-MV3D**, addresses the objective in Section 3.1 through a map-grounded multi-view reasoning pipeline for MLLMs. As illustrated in Figure 3, it decomposes trajectory generation into four stages:

1. Global allocentric map construction from multi-view observations,
2. Question-conditioned view trajectory planning for evidence acquisition,
3. Egocentric grounding of selected evidence for answer prediction,
4. Multi-level dense reward supervision for policy optimization.

This subsection summarizes the pipeline; detailed formulations are presented in Sections 3.3 to 3.6. We denote by  $\pi_{\theta}$  the MLLM-induced policy over the full reasoning trajectory, including intermediate cognitive map generation, view selection, egocentric grounding, and answer prediction.



**Fig. 3: Overall framework with VFM-guided dense rewards.** Given multi-view observations and a question, the MLLM first constructs a *global allocentric cognitive map* and plans a question-conditioned *view trajectory* for evidence acquisition, then grounds the selected views into *egocentric cognitive maps* to predict the final answer. During GRPO optimization, we provide multi-level verifiable rewards: *Global Reward* regularizes the MLLM allocentric map toward a geometry-consistent pseudo target produced by frozen VFMs (e.g., VGGT+SAM3); *Local Reward* supervises the step-wise view trajectory by matching planned viewpoints to a reference trajectory; *Answer Reward* scores answer correctness; and *Format Reward* enforces a valid output format.

### 3.3 Global Allocentric Cognitive Map

Motivated by prior work [15, 44, 52], we construct a global allocentric cognitive map that summarizes scene-level geometry. An allocentric representation facilitates a stable, world-centered coordinate system that remains invariant to camera movement, thereby maintaining geometric consistency across disparate viewpoints. Formally, the global map is defined as:

$$\mathcal{C}^{\text{allo}} \sim \pi_{\theta}(\cdot | \mathcal{I}). \quad (2)$$

where  $\pi_{\theta}$  aggregates multi-view observations into a scene-level representation. As an auxiliary geometric prior, we leverage a pretrained and frozen 3D Vision Foundation Model (VFM), instantiated as a combination of VGGT and SAM3 [3, 43], to produce a geometry-consistent pseudo ground-truth cognitive map,

$$\mathcal{C}^* = \text{VFMs}(\mathcal{I}) \text{ (e.g., SAM3(VGGT}(\mathcal{I}))), \quad (3)$$

and supervise the predicted allocentric cognitive map via structural alignment:

$$R_{\text{global}} = \text{sim}(\mathcal{C}^{\text{allo}}, \mathcal{C}^*). \quad (4)$$

This supervision anchors the learned map to a physically plausible 3D structure without requiring manual annotations. Further details on the pseudo ground-truth cognitive map construction and the calculation of the similarity function are provided in the supplementary material.

### 3.4 View-Trajectory Reasoning

Given the allocentric cognitive map  $\mathcal{C}^{\text{allo}}$  and the question  $q$ , the policy generates an ordered local trajectory for evidence acquisition. Conditioned on  $q$ , the model prioritizes informative viewpoints while filtering irrelevant visual noise, enabling efficient reasoning in complex 3D scenes. Formally, the view trajectory is defined as  $\mathcal{V} = (v_1, \dots, v_T)$ , where each viewpoint is sampled autoregressively as

$$v_t \sim \pi_\theta(\cdot \mid v_{<t}, \mathcal{C}^{\text{allo}}, q). \quad (5)$$

This stage performs local view selection grounded in the global allocentric map. To quantify question-conditioned trajectory quality, we define a step-wise matching reward against a reference trajectory  $\mathcal{V}^* = (v_1^*, \dots, v_T^*)$ :

$$R_{\text{local}} = \frac{1}{T} \sum_{t=1}^T \mathbb{I}[v_t = v_t^*]. \quad (6)$$

$R_{\text{local}}$  measures the fraction of viewpoints that match the reference trajectory at the corresponding step, providing direct supervision for ordered view planning. The reference trajectory  $\mathcal{V}^*$  is deterministically derived from benchmark metadata by identifying the viewpoint that best observes the queried anchor and target objects.

### 3.5 Egocentric Grounding and Answer Prediction

After local trajectory planning, the model converts global evidence into egocentric representations aligned with the selected views:

$$\mathcal{C}_t^{\text{ego}} \sim \pi_\theta(\cdot \mid \mathcal{C}^{\text{allo}}, v_t, \mathcal{I}, q), \quad (7)$$

where  $\pi_\theta(\cdot)$  performs coordinate rotation, spatial filtering, and reference-frame alignment. This conversion is beneficial because MLLMs are primarily pretrained on egocentric (view-aligned) data, and prior work has shown that perspective alignment plays a critical role in visual-spatial problem solving and reasoning [15, 24]. Translating abstract 3D geometry back into the model’s native representation, therefore, maximizes reasoning performance.

This conversion ensures that local trajectory planning precedes egocentric grounding and yields the sequence  $\mathcal{C}_{1:T}^{\text{ego}}$ . The final answer is predicted from trajectory-conditioned egocentric evidence:

$$\hat{a} = \arg \max_a \pi_\theta(a \mid \mathcal{V}, \mathcal{C}_{1:T}^{\text{ego}}, \mathcal{I}, q). \quad (8)$$

### 3.6 Reward Design and Policy Optimization

The task objective in Section 3.1 is answer-level, but optimizing only  $R_{\text{ans}}$  yields a sparse learning signal for multi-step spatial reasoning. Following recent advances in dense reward learning for multi-step reasoning and vision-language policy optimization [4, 38, 58], we therefore optimize the policy with a trajectory-level *dense reward* that provides intermediate supervision aligned with the reasoning stages in Sections 3.3 to 3.5.

**Dense Reward Design.** The reward for trajectory  $\tau$  is defined as:

$$R(\tau) = \lambda_g \cdot R_{\text{global}} + \lambda_l \cdot R_{\text{local}} + \lambda_a \cdot R_{\text{ans}} + \lambda_f \cdot R_{\text{fmt}}. \quad (9)$$

Each term supervises a different stage of the pipeline. We reuse  $R_{\text{global}}$  from Section 3.3 and  $R_{\text{local}}$  from Section 3.4; the answer-level terms are defined as:

$$R_{\text{ans}} = \mathbb{I}[\hat{a} = a^*], \quad R_{\text{fmt}} = \mathbb{I}[\hat{a} \in \mathcal{A}_{\text{valid}}]. \quad (10)$$

Here,  $R_{\text{global}}$  and  $R_{\text{local}}$  provide intermediate supervision for map quality and trajectory quality, while  $(R_{\text{ans}}, R_{\text{fmt}})$  supervise answer correctness and valid output format. In particular,  $R_{\text{ans}}$  is the terminal component that directly matches the answer-level task objective.

**Policy Optimization with GRPO.** We optimize  $\pi_\theta$  using Group Relative Policy Optimization (GRPO) [14, 38]. For each input  $(\mathcal{I}, q)$ , we sample a group of  $K$  trajectories  $\{\tau_i\}_{i=1}^K$  using the rollout policy  $\pi_{\theta_{\text{old}}}$  and compute dense returns  $R(\tau_i)$ . We form a standardized group-relative advantage:

$$\bar{R} = \frac{1}{K} \sum_{j=1}^K R(\tau_j), \quad A_i = \frac{R(\tau_i) - \bar{R}}{\sqrt{\frac{1}{K} \sum_{j=1}^K (R(\tau_j) - \bar{R})^2 + \xi}}, \quad (11)$$

where  $\xi$  is a small constant for numerical stability. Let  $r_i(\theta) = \frac{\pi_\theta(\tau_i | \mathcal{I}, q)}{\pi_{\theta_{\text{old}}}(\tau_i | \mathcal{I}, q)}$  denote the probability ratio between the current and rollout policies. The GRPO objective is:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(\mathcal{I}, q) \sim \mathcal{D}, \tau_{1:K} \sim \pi_{\theta_{\text{old}}}} & \left[ \frac{1}{K} \sum_{i=1}^K \min \left( r_i(\theta) A_i, \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon) A_i \right) \right. \\ & \left. - \beta D_{\text{KL}}(\pi_\theta(\cdot | \mathcal{I}, q) \| \pi_{\text{ref}}(\cdot | \mathcal{I}, q)) \right]. \end{aligned} \quad (12)$$

We update  $\theta$  by maximizing  $\mathcal{J}_{\text{GRPO}}(\theta)$  via gradient ascent. This objective stabilizes policy updates without requiring a separate value network and improves credit assignment across global mapping, view planning, and answer prediction.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We evaluate on three benchmarks: MindCube [44], VSI-Bench [52], and BLINK (MV) [12]. MindCube tests view-dependent 3D spatial reasoning and directional consistency. The dataset consists of 10K training examples and a 1K evaluation set (MindCube-Tiny); throughout the Experiments section, we denote

MindCube-Tiny as “MindCube” for simplicity. VSI-Bench focuses on embodied route planning and compositional spatial relations. BLINK (MV) evaluates multi-view multimodal reasoning under partial and distributed visual evidence. Additional dataset details (data format, question types, and preprocessing) are provided in the supplementary material.

**Baselines.** We compare against representative multi-image MLLM baselines [6, 7, 20, 36, 40, 51, 61] and prior 3D-aware approaches [2, 44, 50, 59]. We keep baseline discussion brief in the main paper and defer prompts, training details, and additional comparisons to the supplementary material.

**Evaluation Metrics.** We follow standard benchmark protocols: accuracy for MindCube and BLINK (MV), and per-category scores with an overall average (Avg) for VSI-Bench.

**Implementation Details.** DR-MV3D is built on Qwen2.5-VL-3B-Instruct [36]. We first perform supervised fine-tuning (SFT) to learn the structured 3D reasoning pipeline, and then further optimize the model with GRPO using the proposed dense reward. In addition, we construct a pseudo-allocentric cognitive map using VGGT [43] and SAM3 [3], which provides global spatial context for downstream reasoning. Detailed training configurations (data formatting, optimization hyperparameters, GRPO settings, and map construction details) are provided in the supplementary material.

## 4.2 Performance Evaluation

**MindCube.** We compare DR-MV3D with prior approaches on the MindCube benchmark. Results in Table 1 show that our model achieves the best overall accuracy among open-weight multi-image models. The vanilla Qwen2.5-VL-3B-Instruct baseline scores 37.8 overall, indicating that direct prompting alone is insufficient for reliable multi-view spatial reasoning on this benchmark. With supervised fine-tuning (SFT), the overall accuracy increases to 62.4, and dense-reward GRPO further improves it to 66.5. The improvement is most pronounced on *Among* and *Around*, which require compositional reasoning over multiple views and consistent interpretation of relative relations. Our GRPO model reaches 71.3 on *Among* and 73.6 on *Around*, substantially improving both categories compared to the instruct baseline. Moreover, our approach surpasses prior map-based baselines trained on the same backbone family, including MindCube-CGMap-SFT (54.4 overall) and MindCube-CGMap-FFR-RL (53.7 overall). This suggests that dense intermediate supervision provides a more effective learning signal for multi-step spatial reasoning than answer-only optimization, and it improves the model’s ability to integrate evidence across views. This behavior is illustrated in Figures 4 and 5, where the baseline misinterprets ego-directions and selects incorrect viewpoints, whereas our model selects an informative viewpoint based on the inferred egocentric map and arrives at the correct answer via

**Table 1: Results on MindCube (MindCube-Tiny).** We report accuracy (%) on the three question categories (*Rotation, Among, Around*) and the overall score, comparing against proprietary and open-weight multi-image MLLMs as well as prior 3D-aware methods. Our model, built on Qwen2.5-VL-3B, consistently improves with SFT and further gains from GRPO with dense rewards, achieving the best overall performance among open-weight baselines. Anno. indicates whether ground-truth intermediate cognitive-map/trajectory annotations are used during training.

Method	Params	Anno.	MindCube-Tiny			
			Rotation	Among	Around	Overall
Random (chance)	-	✗	38.0	37.0	32.8	34.9
Random (frequency)	-	✗	35.7	33.3	31.8	33.0
<i>Proprietary Models</i>						
GPT-4.1 [20]	-	✗	60.0	46.7	55.0	49.6
Gemini-2.5-Pro [6]	-	✗	85.0	49.2	58.3	59.3
GPT-5 [40]	-	✗	94.5	38.2	68.4	56.3
<i>Open-Weight Multi Image Models</i>						
Bagel-MoT [7]	7B	✗	34.5	31.3	42.8	34.7
Qwen2.5-VL-Instruct [36]	3B	✗	34.0	36.0	45.2	37.8
Qwen2.5-VL-Instruct [36]	7B	✗	37.5	31.2	38.0	34.0
Qwen3-VL-Instruct [51]	4B	✗	26.0	26.7	46.0	31.1
Qwen3-VL-Thinking [51]	4B	✗	55.0	28.7	44.8	37.5
InternVL3 [61]	2B	✗	28.9	36.9	45.6	37.5
InternVL3 [61]	8B	✗	36.5	38.1	53.6	41.5
Spatial-MLLM [50]	4B	✗	32.5	47.5	35.0	42.8
MindCube-RawQA-SFT <sub>(Qwen2.5-VL-Instruct)</sub> [44]	3B	✓	34.0	51.0	67.6	51.7
MindCube-CGMap-SFT <sub>(Qwen2.5-VL-Instruct)</sub> [44]	3B	✓	34.5	54.2	70.8	54.4
MindCube-CGMap-FFR-SFT <sub>(Qwen2.5-VL-Instruct)</sub> [44]	3B	✓	31.5	49.8	65.6	50.1
MindCube-CGMap-FFR-RL <sub>(Qwen2.5-VL-Instruct)</sub> [44]	3B	✓	33.0	53.7	70.4	53.7
SenseNova-SI <sub>(Bagel-MoT)</sub> [2]	7B	✗	37.5	57.1	46.8	50.8
SenseNova-SI <sub>(InternVL3)</sub> [2]	2B	✗	33.5	44.4	40.0	41.2
Think3D <sub>(Qwen3-VL-RL)</sub> [59]	4B	✗	41.7	35.0	39.2	41.7
DR-MV3D (Ours) <sub>(Qwen2.5-VL-Instruct w/ SFT)</sub> w/o Anno.	3B	✗	36.5	52.2	64.0	53.6
DR-MV3D (Ours) <sub>(Qwen2.5-VL-Instruct w/ SFT)</sub>	3B	✓	42.0	66.3	69.2	62.4
DR-MV3D (Ours) <sub>(Qwen2.5-VL-Instruct w/ SFT + GRPO)</sub> w/o Anno.	3B	✗	36.5	60.7	67.6	57.7
DR-MV3D (Ours) <sub>(Qwen2.5-VL-Instruct w/ SFT + GRPO)</sub>	3B	✓	43.0	<b>71.3</b>	<b>73.6</b>	<b>66.5</b>

step-wise reasoning. Additional qualitative examples and analyses are provided in the supplementary material.

**VSI-Bench.** We evaluate sequential route planning from multi-view observations and report four criteria: Route Plan (route planning), Rel. Dir. (relative direction), Rel. Dist. (relative distance), and App. Order (object appearance order). As shown in Table 2, our GRPO model achieves the best average score (37.1) among the compared methods. Our model also attains the comparable results on *Route Plan* (32.4), *Rel. Dir.* (46.6), and *Rel. Dist.* (37.8), which directly measures planning quality and directional reasoning across a sequence of observations. These gains suggest that dense-reward training transfers beyond single-turn QA and improves navigation-style evaluation, and this requires maintaining a consistent egocentric frame over time. Relative to the Qwen2.5-VL-3B-Instruct baseline with 30.4, SFT improves the average score to 34.2, and GRPO provides an additional gain to 37.1 while retaining a compact 3B parameter budget. Overall, the results indicate that our training strategy generalizes to sequential decision-style reasoning from multiple frames beyond static multi-view question answering.

**Table 2: VSI-Bench results.** Comparison with multi-image MLLMs and 3D-aware baselines.

Method	Route Plan	Rel. Dir.	Rel. Dist.	App.	Order Avg
LLaVA-OneVision-0.5B [26]	<b>34.5</b>	36.9	28.9	5.8	26.5
LLaVA-OneVision-7B [26]	29.4	35.2	<b>42.5</b>	24.4	32.9
VILA-1.5-8B [27]	31.0	34.8	32.1	24.8	30.7
InternVL2-2B [5]	30.4	44.1	32.1	6.3	28.2
InternVL2-8B [5]	28.9	33.4	38.0	<b>46.4</b>	36.7
Qwen2.5-VL-3B-Instruct [36]	27.3	43.8	25.9	24.4	30.4
DR-MV3D (Ours) <sub>(Qwen2.5-VL-3B-Instruct w/ SFT)</sub>	30.9	43.9	35.1	26.9	34.2
DR-MV3D (Ours) <sub>(Qwen2.5-VL-3B-Instruct w/ SFT + GRPO)</sub>	32.4	<b>46.6</b>	37.8	31.4	<b>37.1</b>

**Table 3: Results on BLINK (MV).**

Method	BLINK (MV)
<i>Proprietary Models</i>	
GLM-4.5V [42]	39.9
Doubao-1.5 [37]	50.9
GPT-4.1 [20]	36.8
Gemini-2.5-Pro [6]	44.9
<i>Open-Weight Models</i>	
RoBoBrain [21]	55.6
Spatial-MLLM [50]	56.0
VLM-3R [11]	41.4
REVPT [60]	51.9
Qwen2.5-VL-3B [36]	42.1
Qwen3-VL-4B [51]	47.9
Qwen3-VL-4B-GRPO [51]	52.4
Qwen3-VL-4B-T3RL [51]	46.1
Think3D <sub>(Qwen3-VL-4B)</sub> [59]	48.6
Think3D <sub>(Qwen3-VL-4B-T3RL)</sub> [59]	53.4
DR-MV3D (Ours) <sub>(Qwen2.5-VL-3B-Instruct w/ SFT)</sub>	54.9
DR-MV3D (Ours) <sub>(Qwen2.5-VL-3B-Instruct w/ SFT + GRPO)</sub>	<b>56.4</b>

**Table 4: Ablation of components in SFT.**

Allo. Map	Traj.	Ego. Map	Overall
<i>w/o SFT (Vanilla Qwen2.5-VL-3B-Instruct)</i>			
✗	✗	✗	37.8
<i>w/ SFT</i>			
✓	✗	✗	52.8
✗	✓	✗	58.2
✗	✗	✓	52.2
✓	✗	✓	53.6
✓	✓	✓	<b>62.4</b>

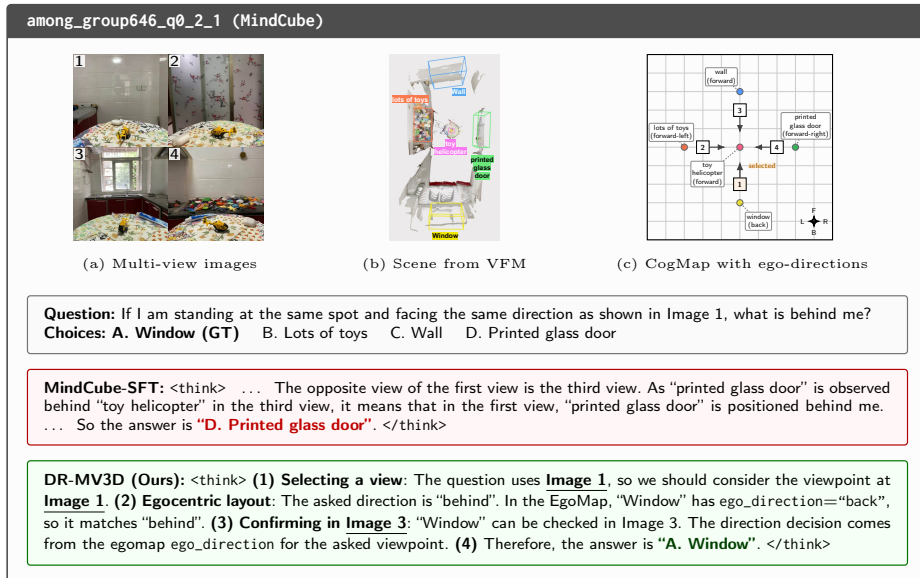
**Table 5: Ablation on rewards in GRPO.**

Answer & format	Global	Local	Overall
✓	✗	✗	63.8
✓	✓	✗	64.9
✓	✓	✓	<b>66.5</b>

**BLINK (MV).** We further evaluate transfer to BLINK (MV), which tests distributed multi-view multimodal reasoning. As shown in Table 3, the Qwen2.5-VL-3B baseline scores 42.1, while our SFT model improves to 54.9, and GRPO further increases it to 56.4. Our GRPO model outperforms large models such as RoBoBrain (55.6) and matches or slightly exceeds Spatial-MLLM (56.0), while also improving over prior RL variants, including Think3D<sub>(Qwen3-VL-4B-T3RL)</sub> (53.4). Significantly, these gains are obtained with the same 3B backbone, which suggests that the improvement comes from the proposed training signal rather than model scaling. These outcomes indicate that dense intermediate supervision generalizes beyond MindCube-style evaluation and improves multi-view reasoning when the answer requires fusing evidence across distributed views.

### 4.3 Ablation Study

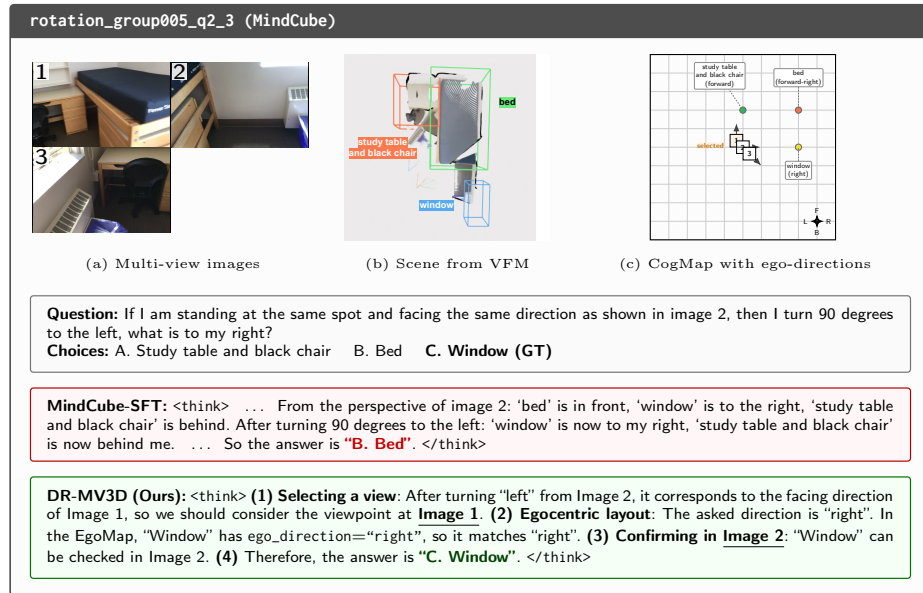
We conduct ablations on MindCube-Tiny to analyze the contribution of (i) the supervision components used in SFT and (ii) the reward design used in GRPO.



**Fig. 4: Qualitative comparison (Among).** Multi-view observations and cognitive-map-based reasoning for an *Among* query (MindCube). The baseline confuses egocentric directions and predicts an incorrect object, whereas our method follows the ego-direction cues in the predicted cognitive map step by step and selects the correct answer.

**Component ablation in SFT.** As shown in Table 4, intermediate supervision is most effective during SFT when multiple components are combined. The vanilla Qwen2.5-3B baseline without SFT achieves 37.8 overall. When we enable a single component, supervising the allocentric map reaches 52.8, supervising the egocentric map reaches 52.2, and supervising only the trajectory reaches 58.2. Supervising both maps improves the score to 53.6, suggesting that the two map representations provide complementary supervision but are not sufficient on their own. Furthermore, adding trajectory supervision for local-view selection yields the best performance of 62.4. This ablation indicates that strong performance benefits from jointly constraining global spatial structure and step-wise reasoning signals, rather than relying on a single intermediate target.

**Reward ablation in GRPO.** We also study the effect of reward design during GRPO, as summarized in Table 5, starting from an SFT model that already produces valid answers and formats. Using only the answer-and-format reward yields 63.8 overall. Adding the global reward improves performance to 64.9, and incorporating both global and local rewards achieves the best result of 66.5. These consistent gains suggest that answer-level optimization alone does not fully capture multi-step spatial reasoning quality, while dense intermediate rewards provide complementary improvements.



**Fig. 5: Qualitative comparison (Rotation).** Multi-view observations and cognitive-map-based reasoning for a *Rotation* query (MindCube). The baseline fails to track the rotated egocentric frame and selects the wrong object, whereas our method maintains consistent ego-direction reasoning under rotation and predicts the correct answer.

## 5 Conclusion

We presented **DR-MV3D**, a dense-reward learning framework for multi-view 3D visual question answering (MV3D-VQA) under partial observability. Our method formulates MV3D-VQA as trajectory-level policy optimization: the agent constructs an allocentric cognitive map to maintain global scene coherence, selects question-conditioned viewpoint trajectories to acquire informative evidence, and aligns observations into egocentric maps for step-wise reasoning and answer prediction. To make these intermediate processes learnable, we introduce dense rewards that directly supervise cross-view consistency and viewpoint selection using pseudo-structural signals from vision foundation models. Experiments on MindCube, VSI-Bench, and BLINK show consistent gains over strong supervised baselines, highlighting the value of process-level supervision for 3D reasoning.

## References

- [1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015) **6**
- [2] 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) **10, 11**
- [3] Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) **4, 7, 10, 5, 8**
- [4] Chan, A.J., Sun, H., Holt, S., Van Der Schaar, M.: Dense reward for free in reinforcement learning from human feedback. arXiv preprint arXiv:2402.00782 (2024) **3, 8**
- [5] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling (2025), <https://arxiv.org/abs/2412.05271> **12**
- [6] Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) **1, 10, 11, 12**
- [7] Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) **10, 11**
- [8] Deng, H., Zou, D., Ma, R., Luo, H., Cao, Y., Kang, Y.: Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. arXiv preprint arXiv:2503.07065 (2025) **6**
- [9] Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv e-prints pp. arXiv–2503 (2025) **6**
- [10] Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023) **2**
- [11] Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., Xu, H., Theiss, J., Chen, T., Li, J., Tu, Z., Wang, Z., Ranjan, R.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction (2025), <https://arxiv.org/abs/2505.20279> **12**

- [12] Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024) [2](#), [5](#), [6](#), [9](#), [7](#)
- [13] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017) [6](#)
- [14] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) [1](#), [5](#), [9](#)
- [15] Guo, Z., Yang, Z., Li, Y., Zhang, X., Gao, W., Wang, J., Li, C., Liu, X., Jian, P.: Can llms see without pixels? benchmarking spatial intelligence from textual descriptions. arXiv preprint arXiv:2601.03590 (2026) [3](#), [5](#), [7](#), [8](#)
- [16] Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2616–2625 (2017) [2](#)
- [17] Hong, Y., Lin, C., Du, Y., Chen, Z., Tenenbaum, J.B., Gan, C.: 3d concept learning and reasoning from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9202–9212 (2023) [1](#), [2](#), [3](#), [5](#), [6](#)
- [18] Huang, T., Zhang, Z., Tang, H.: 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. arXiv preprint arXiv:2507.23478 (2025) [2](#), [5](#)
- [19] Huang, X., Wu, J., Xie, Q., Han, K.: 3drs: Mllms need 3d-aware representation supervision for scene understanding. arXiv preprint arXiv:2506.01946 (2025) [5](#)
- [20] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) [1](#), [10](#), [11](#), [12](#)
- [21] Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., Xue, X., Su, Q., Lyu, H., Zheng, X., Liu, J., Wang, Z., Zhang, S.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete (2025), <https://arxiv.org/abs/2502.21257> [12](#)
- [22] 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) [2](#)
- [23] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9579–9589 (2024) [6](#)
- [24] Lee, P.Y., Je, J., Park, C., Uy, M.A., Guibas, L., Sung, M.: 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) 5, 8
- [25] Lee, S., Choi, J., Kang, I., Kim, J., Park, J., Shim, H.: 3d-aware vision-language models fine-tuning with geometric distillation. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 10628–10647 (2025) 5
- [26] Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), <https://arxiv.org/abs/2408.03326> 12
- [27] Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models (2024), <https://arxiv.org/abs/2312.07533> 12
- [28] Linghu, X., Huang, J., Niu, X., Ma, X.S., Jia, B., Huang, S.: Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems* **37**, 140903–140936 (2024) 5
- [29] Linghu, X., Huang, J., Zhu, Z., Jia, B., Huang, S.: Scenecot: Eliciting grounded chain-of-thought reasoning in 3d scenes. *arXiv preprint arXiv:2510.16714* (2025) 5
- [30] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023) 1
- [31] Liu, Q., Huang, T., Zhang, Z., Tang, H.: Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884* (2025) 5
- [32] Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520* (2025) 6
- [33] Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. *arXiv preprint arXiv:2505.12081* (2025) 6
- [34] Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2034–2044 (2025) 5
- [35] Peng, Y., Zhang, G., Zhang, M., You, Z., Liu, J., Zhu, Q., Yang, K., Xu, X., Geng, X., Yang, X.: Lmm-r1: Empowering 3b lmm with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536* (2025) 6
- [36] Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), <https://arxiv.org/abs/2412.15115> 1, 3, 5, 10, 11, 12, 4
- [37] Seed, B., :, Chen, J., Fan, T., Liu, X., Liu, L., Lin, Z., Wang, M., Wang, C., Wei, X., Xu, W., Yuan, Y., Yue, Y., Yan, L., Yu, Q., Zuo, X., Zhang, C., Zhu, R., An, Z., Bai, Z., Bao, Y., Bin, X., Chen, J., Chen, F., Chen,

- H., Chen, R., Chen, L., Chen, Z., Chen, J., Chen, S., Chen, K., Chen, Z., Chen, J., Chen, J., Chi, J., Dai, W., Dai, N., Dai, J., Dou, S., Du, Y., Du, Z., Duan, J., Dun, C., Fan, T.H., Feng, J., Feng, J., Feng, Z., Fu, Y., Fu, W., Fu, H., Ge, H., Guo, H., Han, M., Han, L., Hao, W., Hao, X., He, Q., He, J., He, F., Heng, W., Hong, Z., Hou, Q., Hu, L., Hu, S., Hu, N., Hua, K., Huang, Q., Huang, Z., Huang, H., Huang, Z., Huang, T., Huang, W., Jia, W., Jia, B., Jia, X., Jiang, Y., Jiang, H., Jiang, Z., Jiang, K., Jiang, C., Jiao, J., Jin, X., Jin, X., Lai, X., Li, Z., Li, X., Li, L., Li, H., Li, Z., Wan, S., Wang, Y., Li, Y., Li, C., Li, N., Li, S., Li, X., Li, X., Li, A., Li, Y., Liang, N., Liang, X., Lin, H., Lin, W., Lin, Y., Liu, Z., Liu, G., Liu, G., Liu, C., Liu, Y., Liu, G., Liu, J., Liu, C., Liu, D., Liu, K., Liu, S., Liu, Q., Liu, Y., Liu, K., Liu, G., Liu, B., Long, R., Lou, W., Lou, C., Luo, X., Luo, Y., Lv, C., Lv, H., Ma, B., Ma, Q., Ma, H., Ma, Y., Ma, J., Ma, W., Ma, T., Mao, C., Min, Q., Nan, Z., Ning, G., Ou, J., Pan, H., Pang, R., Peng, Y., Peng, T., Qian, L., Qian, L., Qiao, M., Qu, M., Ren, C., Ren, H., Shan, Y., Shen, W., Shen, K., Shen, K., Sheng, G., Shi, J., Shi, W., Shi, G., Cao, S.S., Song, Y., Song, Z., Su, J., Sun, Y., Sun, T., Sun, Z., Wan, B., Wang, Z., Wang, X., Wang, X., Wang, S., Wang, J., Wang, Q., Wang, C., Wang, S., Wang, Z., Wang, C., Wang, J., Wang, S., Wang, X., Wang, Z., Wang, Y., Wang, W., Wang, T., Wei, C., Wei, H., Wei, Z., Wei, S., Wu, Z., Wu, Y., Wu, Y., Wu, B., Wu, S., Wu, J., Wu, N., Wu, S., Wu, J., Xi, C., Xia, F., Xian, Y., Xiang, L., Xiang, B., Xiao, B., Xiao, Z., Xiao, X., Xiao, Y., Xin, C., Xin, S., Xiong, Y., Xu, J., Xu, Z., Xu, C., Xu, J., Xu, Y., Xu, W., Xu, Y., Xu, S., Yan, S., Yan, S., Yang, Q., Yang, X., Yang, T., Yang, Y., Yang, Y., Yang, X., Yang, Z., Yang, G., Yang, Y., Yao, X., Yi, B., Yin, F., Yin, J., Ying, Z., Yu, X., Yu, H., Yu, S., Yu, M., Yu, H., Yuan, S., Yuan, J., Zeng, Y., Zhan, T., Zhang, Z., Zhang, Y., Zhang, M., Zhang, W., Zhang, R., Zhang, Z., Zhang, T., Zhang, X., Zhang, Z., Zhang, S., Zhang, W., Zhang, X., Zhang, Y., Zhang, Y., Zhang, G., Zhang, H., Zhang, Y., Zheng, R., Zheng, N., Zheng, Z., Zheng, Y., Zheng, C., Zhi, X., Zhong, W., Zhong, C., Zhong, Z., Zhong, B., Zhou, X., Zhou, N., Zhou, H., Zhu, H., Zhu, D., Zhu, W., Zuo, L.: Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning (2025), <https://arxiv.org/abs/2504.13914> **12**
- [38] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) **5, 8, 9**
- [39] Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024) **4**
- [40] Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2026) **10, 11**
- [41] Su, Y., Yu, D., Song, L., Li, J., Mi, H., Tu, Z., Zhang, M., Yu, D.: Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. arXiv preprint arXiv:2503.23829 (2025) **5**

- [42] Team, V., 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., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Li, H., Zhu, J., Chen, J., Xu, J., Xu, J., Chen, J., Lin, J., Chen, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Lyu, R., Tu, S., Yang, S., Meng, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Zhang, T., Luo, T., Hao, T., Tong, T., Jia, W., Li, W., Liu, X., Zhang, X., Lyu, X., Zhang, X., Fan, X., Huang, X., Xue, Y., Wang, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Huang, Y., Niu, Y., Shi, Y., Wang, Y., Wang, Y., Yue, Y., Li, Y., Liu, Y., Zhang, Y., Wang, Y., Zhang, Y., Xue, Z., Du, Z., Hou, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., Tang, J.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2026), <https://arxiv.org/abs/2507.01006> **12**
- [43] Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025) **3, 4, 7, 10, 5, 8**
- [44] Wang, Q., Yin, B., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., Xie, S., Li, M., Wu, J., Fei-Fei, L.: Understanding VLMs spatial mental modeling capability from limited views. In: The Fourteenth International Conference on Learning Representations (2026), <https://openreview.net/forum?id=0FhrtdKLTd> **2, 3, 5, 6, 7, 9, 10, 11, 4**
- [45] Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024) **3**
- [46] 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) **1**
- [47] Wang, Z., Li, X., Yang, J., Liu, Y., Jiang, S.: Gridmm: Grid memory map for vision-and-language navigation. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 15625–15636 (2023) **2**
- [48] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022) **6**
- [49] Wen, X., Liu, Z., Zheng, S., Xu, Z., Ye, S., Wu, Z., Liang, X., Wang, Y., Li, J., Miao, Z., et al.: Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. arXiv preprint arXiv:2506.14245 (2025) **5**
- [50] Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mlm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025) **10, 11, 12**

- [51] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), <https://arxiv.org/abs/2505.09388> **1, 10, 11, 12**
- [52] Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025) **2, 5, 7, 9**
- [53] 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) **5, 16**
- [54] Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2376–2385 (2025) **6**
- [55] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022) **6**
- [56] Yuan, Y., Cui, H., Huang, Y., Chen, Y., Ni, F., Dong, Z., Li, P., Zheng, Y., Hao, J.: Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. arXiv preprint arXiv:2508.13998 (2025) **5**
- [57] Zhan, Y., Zhu, Y., Zheng, S., Zhao, H., Yang, F., Tang, M., Wang, J.: Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. arXiv preprint arXiv:2503.18013 (2025) **5**
- [58] Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization (2025), <https://arxiv.org/abs/2503.12937> **3, 5, 8**
- [59] Zhang, Z., Wu, Y., Jia, L., Wang, Y., Zhang, Z., Li, Y., Ran, B., Zhang, F., Sun, Z., Yin, Z., et al.: Think3d: Thinking with space for spatial reasoning. arXiv preprint arXiv:2601.13029 (2026) **1, 2, 3, 5, 10, 11, 12**
- [60] Zhou, Z., Chen, D., Ma, Z., Hu, Z., Fu, M., Wang, S., Wan, Y., Zhao, Z., Krishna, R.: Reinforced visual perception with tools (2025), <https://arxiv.org/abs/2509.01656> **12**
- [61] Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) **10, 11**

- [62] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023) [2](#)

# Dense Reward for Multi-View 3D Reasoning with Global Maps and Local Views

*(Supplementary Material)*

## Supplementary Material

S1 Overview .....	2
S2 Limitations & Future Work .....	2
S3 Method Details .....	2
S3.1 Cognitive Map Representation .....	2
S3.2 Cognitive Map Similarity for Global Reward .....	3
S4 Implementation Details .....	4
S4.1 Input-Output Formatting. ....	5
S4.2 SFT Data Construction. ....	6
S4.3 Pseudo Cognitive Map Construction. ....	8
S4.4 Optimization for Supervised Fine-tuning. ....	8
S4.5 GRPO Training and Reward Implementation. ....	8
S4.6 Inference and Evaluation. ....	9
S4.7 Code Release. ....	9
S5 Dataset and Evaluation Details .....	9
S5.1 MindCube .....	10
S5.2 VSI-Bench .....	12
S5.3 BLINK-MV .....	12
S6 Additional Experimental Results .....	12
S6.1 Additional Qualitative Results .....	12
S6.2 Training Plot and Reward Plot Analysis. ....	14
S6.3 Effects of Egocentric Grounding .....	15
S6.4 Failure Cases .....	16

## S1 Overview

This supplementary material provides additional details on the proposed map-grounded learning framework for multi-view 3D reasoning. In particular, we describe the structured reasoning pipeline, the construction of supervised training data, the implementation and optimization details for SFT and GRPO, and the dataset-specific evaluation protocols used in our experiments. We also include additional analyses and qualitative examples to complement the results in the main paper.

## S2 Limitations & Future Work

While this work focuses on static scenes, extending the framework to dynamic environments (e.g., video, 4D) is challenging due to temporal correspondence, motion, and scene changes. Nevertheless, the proposed formulation can be naturally generalized by defining temporally-aware consistency and exploration rewards. Another promising direction is multi-turn spatial reasoning. Iterative querying with memory updates and long-horizon planning could support more precise, compositional 3D inference and improve robustness in complex scenarios.

## S3 Method Details

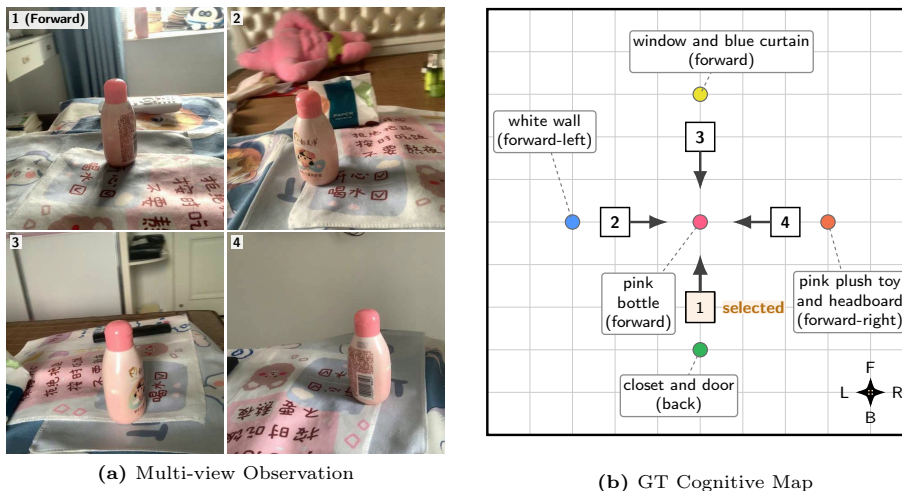
This section provides additional methodological details deferred from the main paper, including the representation of cognitive maps and the similarity function used for global reward computation.

### S3.1 Cognitive Map Representation

Following previous cognitive-map-based formulations for multi-view spatial reasoning [16, 44, 47, 52], we represent the allocentric cognitive map as a structured bird’s-eye-view scene abstraction defined on a discretized  $10 \times 10$  grid [44, 52]. As shown in Figures A1a and A1b, multi-view observations are integrated into a unified allocentric representation, where the resulting cognitive map captures the coarse spatial organization of the scene in a world-centered reference frame.

Each cognitive map consists of two types of entries: *objects*, which denote salient scene entities, and *views*, which denote camera viewpoints associated with the input images. In practice, each cognitive map is serialized as a lightweight JSON structure containing object and viewpoint entries. Each object entry contains its semantic name and grid position, and may additionally include a discrete facing attribute when such orientation is available. Each view entry contains its identifier, grid position, and a discrete facing direction. We use a small set of canonical orientation labels, namely *up*, *down*, *left*, and *right*, to represent viewpoint orientation on the grid.

This representation is intentionally lightweight and structured. Rather than modeling metrically precise 3D geometry, it captures the coarse spatial layout



**Fig. A1:** Example of the allocentric cognitive map representation. **(a)** Multi-view observations of a scene captured from different viewpoints. **(b)** The corresponding ground-truth cognitive map represented on a discretized bird’s-eye-view grid, where objects and viewpoints are placed according to their relative spatial arrangement in a world-centered reference frame.

that is sufficient for downstream trajectory planning and grounded reasoning. In particular, the grid coordinates should be interpreted as a compact structural abstraction of the scene, where object and viewpoint locations indicate their relative arrangement in the global map. The resulting representation is both interpretable and easy for the MLLM to generate as an explicit intermediate state.

Formally, the allocentric cognitive map is represented as:

$$\mathcal{C}^{\text{allo}} = \{\mathcal{O}^{\text{allo}}, \mathcal{V}^{\text{allo}}\}, \quad (\text{A1})$$

where  $\mathcal{O}^{\text{allo}}$  denotes the set of object entries and  $\mathcal{V}^{\text{allo}}$  denotes the set of viewpoint entries. Each object entry is described by its name and position, while each viewpoint entry is described by its name, position, and facing direction. This structured map serves as the global intermediate representation on top of which the subsequent local trajectory planning and egocentric grounding are performed.

### S3.2 Cognitive Map Similarity for Global Reward

Given the predicted allocentric map  $\mathcal{C}^{\text{allo}}$  and the pseudo target  $\mathcal{C}^*$ , the global reward is defined by a structural similarity function,

$$R_{\text{global}} = \text{sim}(\mathcal{C}^{\text{allo}}, \mathcal{C}^*). \quad (\text{A2})$$

The similarity score evaluates whether the predicted map preserves the spatial structure of the reference map. Importantly, the objective does not require exact alignment of absolute coordinates on the discretized  $10 \times 10$  grid. Instead, it emphasizes the *relative* spatial configuration among salient objects in the ground-truth map, since preserving such relational structure is more important for downstream reasoning than reproducing exact positions.

We compute the similarity from two complementary components,

$$\text{sim}(\mathcal{C}^{\text{allo}}, \mathcal{C}^*) = \alpha_{\text{sim}} s_{\text{dir}} + (1 - \alpha_{\text{sim}}) s_{\text{face}}, \quad (\text{A3})$$

where  $s_{\text{dir}} \in [0, 1]$  denotes directional similarity,  $s_{\text{face}} \in [0, 1]$  denotes facing similarity, and  $\alpha_{\text{sim}} \in [0, 1]$  balances the two terms.

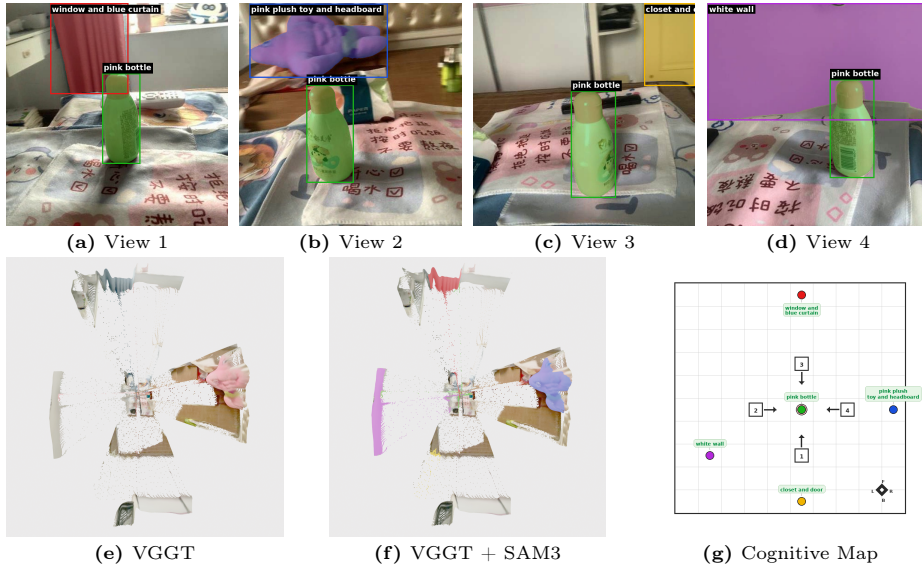
For **directional similarity**, we first identify the set of salient objects shared by  $\mathcal{C}^{\text{allo}}$  and  $\mathcal{C}^*$ , where the reference objects in  $\mathcal{C}^*$  define the comparison set. We then compare the coarse pairwise relation of each object pair in the two maps. Each pair is assigned a discrete relative direction such as *left*, *right*, *up*, *down*, or an inward/outward relation when the pair is spatially coincident under the adopted map convention. The directional score is given by the fraction of object pairs whose relative relations agree between the two maps. This formulation makes the reward invariant to global translation and robust to small coordinate perturbations, as long as the relative layout is preserved.

For **facing similarity**, we compare the facing direction associated with each valid viewpoint in the predicted and reference maps. The score is defined as the fraction of reference viewpoints whose discrete facing labels match those in the prediction. Therefore, this term evaluates whether the model not only captures the spatial layout of the scene, but also preserves the viewpoint orientation needed for grounded reasoning.

Overall, the global similarity function favors structurally consistent cognitive maps rather than surface-level coordinate matching. As a result, the model is rewarded for recovering the relative scene geometry and viewpoint orientation that are most relevant to trajectory planning and 3D reasoning.

## S4 Implementation Details

Our framework is built upon Qwen2.5-VL [36] and is implemented by extending the official MindCube [44] codebases. We use Qwen2.5-VL as the main multi-modal backbone throughout the training pipeline, allowing us to study the effect of structured supervision and dense reward optimization under a fixed model architecture. Unless otherwise noted, we use a maximum sequence length of 8,192 tokens so that the model can process multi-view visual inputs, task instructions, intermediate reasoning fields, and the final answer within a unified generation format. We first train the model with supervised fine-tuning (SFT) on structured targets, and then further optimize it with Group Relative Policy Optimization (GRPO). For GRPO training, we use the ver1 [39] framework, which provides an efficient infrastructure for reward-based optimization over large multimodal policies. In the remainder of this section, we describe the input-output format, the



**Fig. A2: Cognitive map construction from multi-view observations.** We first segment salient objects in the question with SAM3 [3] across multi-view images. Then, we reconstruct the scene with 3D vision foundation model such as VGGT [43], enriching it with semantic grounding from 2D segmentation results from multi-views. The bottom row shows the VGGT reconstruction, the semantically grounded scene after SAM3, and the final cognitive map. The final cognitive map summarizes the spatial layout in an allocentric bird’s-eye view.

SFT data construction and optimization procedure, the pseudo cognitive map construction, the GRPO setup and reward implementation, and the inference details necessary to reproduce our method.

#### S4.1 Input-Output Formatting.

Each training instance consists of a set of multi-view images, a question, and a structured target sequence. We use a unified autoregressive output format with four blocks:

```

<CogMap> ... </CogMap>
  <think> ...
<EgoMap> ... </EgoMap>
  ... </think>
<answer> ... </answer>

```

This format enables the model to jointly learn intermediate spatial representations and final answer prediction within a single generation stream. The <CogMap> block represents the global allocentric scene layout, the <EgoMap> block represents the viewpoint-aligned local layout relevant to the question, the <think>

block verbalizes the grounded reasoning process, and the <answer> block contains the final answer used for evaluation. We use a maximum sequence length of 8,192 tokens so that multi-view visual inputs, instructions, intermediate reasoning fields, and the final answer can be processed within a single context window.

## S4.2 SFT Data Construction.

```

Example of SFT target construction (MindCube)

[Rules]
...
[Question]
...

[Answer]
<CogMap>{
  "objects": [
    {"name": "black waist bag", "position": [5, 5]},
    {"name": "plush toy", "position": [5, 8]},
    {"name": "black sofa", "position": [2, 5], "facing": "down"},
    {"name": "window", "position": [5, 2]},
    {"name": "display shelves", "position": [8, 5]}
  ],
  "views": [
    {"name": "Image 3", "position": [5, 4], "facing": "down"},
    {"name": "Image 1", "position": [5, 6], "facing": "up"},
    {"name": "Image 2", "position": [4, 5], "facing": "right"},
    {"name": "Image 4", "position": [6, 5], "facing": "left"}
  ]
}</CogMap>
<think>
(1) Selecting a view: The question references Image 3 at [5, 4] facing="down". We should consider the viewpoint from Image 3.

(2) Egocentric layout:
<EgoMap viewpoint="Image 3">
- black waist bag: ego_direction="forward"
- plush toy: ego_direction="forward"
- black sofa: ego_direction="forward-right"
- window: ego_direction="back"
- display shelves: ego_direction="forward-left"
</EgoMap>

Answering from the EgoMap: The question asks about "left". In the EgoMap, "Display shelves" has ego_direction="forward-left", so it matches "left".

(3) Confirming in Image 1: Display shelves is visible, consistent with ego_direction="forward-left" in the EgoMap.

(4) Therefore, the answer is "A. Display shelves".</think>
<answer>A. Display shelves.</answer>

```

**Fig. A3: Example of a structured SFT target.** Starting from benchmark meta-data, we programmatically construct a unified supervision target composed of an allocentric cognitive map, egocentric grounding, structured reasoning trace, and final answer. The model is trained to autoregressively generate this full trajectory rather than only the final answer.

Starting from the original benchmark training examples [12, 44, 52], we construct supervised fine-tuning (SFT) targets through a fully programmatic, template-based pipeline following the data construction strategy of MindCube [44]. Rather than relying on LLM-generated rationales, all intermediate supervision signals are produced deterministically from the provided ground-truth metadata using predefined templates and rule-based transformations. Each raw sample consists of a multi-view image set, a question, a ground-truth answer, and scene metadata describing object identities, spatial relations, and viewpoint attributes. Our goal is to convert each sample into a structured reasoning trajectory aligned with the four stages of our framework: allocentric cognitive map construction, egocentric map grounding, grounded reasoning, and final answer prediction.

For each example, we first construct an allocentric cognitive map from the scene metadata and viewpoint annotations. This representation places all objects and camera viewpoints into a canonical bird’s-eye grid in a world-centered coordinate frame, yielding a global scene representation that is consistent across views. Although the exact placement rules vary slightly across benchmark settings, the construction is deterministic in all cases and explicitly encodes object identities, coarse coordinates, and viewpoint poses.

Next, we derive an egocentric map conditioned on the question. Specifically, we parse the question to determine the referenced viewpoint as well as any camera transformation implied by the query, such as a rotation or change of facing direction. Using this parsed view specification, we compute the effective camera pose and convert object locations from the allocentric map into egocentric directions relative to that pose. The resulting representation specifies which objects lie in directions such as front, back, left, right, or diagonal sectors, and also records the view that most directly supports the answer object.

Based on the allocentric and egocentric representations, we then generate a grounded reasoning chain using a fixed template that mirrors the intended allocentric-to-egocentric reasoning flow. The reasoning trace first summarizes the global scene layout and camera poses, then identifies the selected viewpoint and any pose update, next describes the egocentric object arrangement, and finally derives the answer through explicit elimination over candidate objects. To preserve visual grounding, object references in the reasoning trace are annotated with their supporting image view whenever applicable.

Finally, we serialize the constructed supervision into a single structured target sequence consisting of <CogMap>, <EgoMap>, <think>, and <answer>. These components are concatenated so that the model is trained to autoregressively generate the full intermediate reasoning trajectory, rather than only the final answer.

The complete SFT data pipeline consists of scaffold generation, prompt construction, model-specific format conversion, and dataset patching for training. Across all stages, the intermediate representations are generated programmatically from benchmark metadata, ensuring consistency, reproducibility, and close alignment with the template-based supervision protocol introduced in MindCube. Our final SFT corpus contains 10,000 training examples, each associated

with 2 to 6 input images. We also construct a held-out evaluation split from the benchmark tinybench subset containing 1,050 examples. A representative example of the resulting structured SFT target is shown in Figure A3.

### S4.3 Pseudo Cognitive Map Construction.

To provide a geometrically grounded signal for allocentric map generation, we construct a pseudo ground-truth cognitive map  $\mathcal{C}^*$  from the multi-view image set  $\mathcal{I}$  using frozen 3D vision foundation models (VFMs). When ground-truth allocentric maps are available in the dataset, we use them directly; otherwise, we use the VFM-derived pseudo map as the supervision target. Concretely, we first apply a pretrained 3D reconstruction model (e.g., VGGT [43]) to infer a scene-level geometric representation from the input views, and then use a segmentation model (e.g., SAM3 [3]) to identify salient object instances and their spatial extents in the reconstructed scene. The resulting 3D scene is subsequently projected into a structured allocentric representation that summarizes object identities, coarse locations, and spatial relations in a world-centered reference frame. This process yields a geometry-consistent pseudo target:

$$\mathcal{C}^* = \text{VFMs}(\mathcal{I}) = \mathcal{F}(\text{SAM3}(\text{VGGT}(\mathcal{I}))), \quad (\text{A4})$$

where  $\mathcal{F}$  denotes a coarse scene-to-map transformation that converts the reconstructed object-centric 3D scene into the allocentric cognitive-map format used by the language model. In particular,  $\mathcal{F}$  abstracts away fine-grained geometry by discretizing object positions into a canonical world-centered layout and summarizing pairwise spatial relations at a coarse semantic level. The pseudo map is not intended to be a perfect annotation. Rather, it provides a stable and verifiable proxy supervision signal that encourages the predicted allocentric cognitive map to remain consistent with the underlying 3D scene geometry. This design is particularly useful because manually annotating global cognitive maps for large-scale multi-view reasoning data is expensive and difficult to standardize.

### S4.4 Optimization for Supervised Fine-tuning.

During supervised fine-tuning (SFT), we optimize the model using the standard next-token cross-entropy objective over the full target sequence. This trains the model to autoregressively generate the complete reasoning trajectory, including intermediate map tokens and the final answer. Unless otherwise noted, we use a global batch size of 256, a per-device batch size of 1, a learning rate of  $1 \times 10^{-5}$ , weight decay of 0, and a warmup ratio of 0.03, and train for 3 epochs. We use AdamW as the optimizer with a cosine learning rate schedule. Training is conducted on 8 NVIDIA H100 GPUs.

### S4.5 GRPO Training and Reward Implementation.

After supervised fine-tuning, we further optimize the policy using GRPO with the dense trajectory-level reward defined in Equation (9) of the main paper:

$$R(\tau) = \lambda_g R_{\text{global}} + \lambda_l R_{\text{local}} + \lambda_a R_{\text{ans}} + \lambda_f R_{\text{fmt}}.$$

Here, the global reward measures the structural similarity between the predicted allocentric cognitive map and the target cognitive map, while the local reward compares the predicted ordered view trajectory against a reference trajectory and computes the fraction of step-wise matches. The answer reward is given by exact match on the final answer, and the format reward checks whether the generated output follows the required structured format and contains a valid answer field. In practice, the format reward helps suppress malformed generations during reinforcement learning, whereas the global and local rewards provide intermediate supervision signals beyond answer-only optimization.

For each input, we sample a group of rollouts from the current policy and compute group-relative standardized advantages following Equation (11). We then optimize the clipped GRPO objective in Equation (12), together with a KL regularization term that constrains the learned policy toward a reference model. In our implementation, the SFT checkpoint is used both to initialize the policy and as the KL reference policy, and GRPO is implemented using the `verl` framework. Unless otherwise noted, we use a rollout group size of  $K = 8$ , clipping coefficient  $\epsilon = 0.2$ , KL coefficient  $\beta = 0.001$ , and a learning rate of  $1 \times 10^{-6}$ , and train for 600 update steps. The reward weights are set to  $\lambda_g = 0.5$ ,  $\lambda_l = 0.2$ ,  $\lambda_a = 0.5$ , and  $\lambda_f = 0.2$ .

#### S4.6 Inference and Evaluation.

At inference time, the model receives the same multi-view input format as in training and produces a single structured response. We extract the final prediction from the `<answer>` field for benchmark evaluation. For MindCube and BLINK (MV), we report accuracy. For VSI-Bench, we report category-wise scores and the overall average following the benchmark protocol. For VSI-Bench in particular, although the model is trained on static multi-view images, we adapt the inference prompt to sampled video frames by instructing the model to first build an egocentric map from the frames and then follow the same coordinate-based reasoning pipeline used during training.

#### S4.7 Code Release.

We will release the training and inference code, together with data preprocessing scripts and evaluation utilities.

## S5 Dataset and Evaluation Details

This section summarizes the benchmark-specific input setting and evaluation protocol used in our experiments. We evaluate on three benchmarks: MindCube, VSI-Bench, and BLINK (MV). Following the benchmark conventions, we report accuracy on MindCube and BLINK (MV), and category-wise scores together with the overall average on VSI-Bench.

**Prompt for MindCube dataset**

**[System Prompt]**  
 You are a spatial reasoning assistant for multi-view 3D scenes. Given multiple images showing a 3D scene from different viewpoints, reason carefully about the spatial layout with constructing an allocentric cognitive map (CogMap) and answer the question. The CogMap is a structured grid where each object has an (x, y) grid coordinate in the allocentric (world-centered) frame, and each viewpoint has a position, facing direction, and associated image.

**[Answer Instruction]**  
 Output 3 blocks in order:

1. <CogMap>: Construct an allocentric cognitive map from multi-view images.
2. <think>: Complete a structured reasoning trajectory with 4 numbered steps:
  - (1) Selecting a view: Read the question, identify the spatial anchor, locate it in the CogMap, and determine which viewpoint is most relevant, e.g., We should consider the viewpoint from <Image X>.
  - (2) Egocentric layout: Convert the allocentric CogMap into an egocentric frame aligned with the selected viewpoint's facing direction. Output as:
 

```
<EgoMap viewpoint="Image X">
- [object_name]: ego_direction="[direction]"
</EgoMap>
```

If rotation/movement is required, apply it and describe updated directions. Then, map the question's spatial query to ego-direction labels, e.g., A.name=ego\_dir [Image X].
  - (3) Confirmation: Reference another image that visually confirms the answer, e.g., Confirming in <Image Y>: [object] is visible, consistent with the EgoMap.
  - (4) State the final answer: Therefore, the answer is "X. option\_text".
3. <answer>: "X. option\_text"

**[CogMap Instruction]**  
 <cogmap\_gen\_instruction>

**[Egocentric Map Rules]**

- The viewpoint's facing direction becomes "forward" in the EgoMap.
- Objects in the opposite direction become "back".
- Apply 90° rotation for left/right accordingly.
- For diagonal facings, decompose and rotate consistently.
- Include only objects that are spatially relevant from the selected viewpoint.
- Rotation rules: Grid [0,0]=top-left, [9,9]=bottom-right.
  - facing=up → forward=[-y], right=[+x]
  - facing=right → forward=[+x], right=[+y]
  - facing=down → forward=[+y], right=[-x]
  - facing=left → forward=[-x], right=[-y]

**Fig. A4: Prompt template for MindCube-Tiny.** We follow the prompts of MindCube for allocentric cognitive map generation, and add explicit egocentric representations.

## S5.1 MindCube

MindCube evaluates spatial reasoning under partial observations and dynamic viewpoints using multi-view image groups paired with multiple-choice spatial questions. The benchmark is organized around three camera movement settings: *Rotation*, *Among*, and *Around*, which respectively test in-place view rotation,

**Prompt for VSI-Bench evaluation**

**[System Prompt]**  
 You are a spatial reasoning assistant for 3D scenes captured as a multi-frame video. Given frames uniformly sampled from a video of a 3D scene, construct an allocentric cognitive map (CogMap) from the visual observations and reason about the spatial layout to answer the question.

**[Answer Instruction]**  
 Output 3 blocks in order:

1. <CogMap>: Construct an allocentric cognitive map as a  $10 \times 10$  grid by observing the video frames. Assign (x, y) coordinates to each object based on their estimated positions in the scene.
2. <think>: Complete a structured reasoning trajectory with 4 numbered steps:
  - (1) Selecting a viewpoint: Identify the standing position and facing target in the CogMap from the question, and compute the facing direction.
  - (2) Egocentric layout: Convert the allocentric CogMap into an egocentric frame aligned with the standing position’s facing direction. Output as:
 

```
<EgoMap position="[object_name]">
- [object_name]: ego_direction="[direction]"
</EgoMap>
```

From the standing position, compute each direction to relevant objects.
  - (3) Confirmation: Reference the video frames to visually confirm the spatial relationship.
  - (4) State the final answer: Therefore, the answer is “X. option\_text”.
3. <answer>: “X. option\_text”.

**[CogMap Instruction]**  
 <CogMap\_gen\_instruction>

**[Egocentric Map Rules]**

- The facing direction (from standing position toward the target object) becomes “forward” in the ego frame.
- Compute displacement vectors from the standing position to each queried object.
- Rotation rules: Grid [0,0]=top-left, [9,9]=bottom-right.
  - facing=up → forward=[-y], right=[+x]
  - facing=right → forward=[+x], right=[+y]
  - facing=down → forward=[+y], right=[-x]
  - facing=left → forward=[-x], right=[-y]

**Fig. A5: Prompt template for VSI-Bench evaluation.** Since VSI-Bench provides multi-frame video inputs with object-centric, position-based questions with objects rather than explicit camera viewpoints, we adapt the MindCube-Tiny prompt accordingly.

movement among objects, and movement around the scene. The questions cover viewpoint-dependent spatial relations, perspective taking, and hypothetical motion, and are designed to require reasoning about objects that may not be visible in the current view. In our experiments, we follow the MindCube-Tiny protocol used in the main paper, which contains 10K training examples and a 1K evaluation split; for brevity, we denote this evaluation set as *MindCube* throughout the

paper. We use the benchmark-specific prompting template shown in Figure A4, and report accuracy on each of the three settings and the overall average.

### S5.2 VSI-Bench

VSI-Bench is a video-based visual-spatial intelligence benchmark built from real indoor-scene videos with object-level 3D annotations, covering a diverse set of spatial reasoning tasks such as object count, relative distance, relative direction, route planning, object size, absolute distance, room size, and appearance order. Unlike MindCube, VSI-Bench poses questions grounded in physical positions and orientations in the scene (e.g., “standing by the door and facing the bookshelf”) rather than referencing specific camera views. In addition, the visual input is given as uniformly sampled video frames instead of discrete multi-view images with known viewpoints. To bridge this domain gap, we adapt the inference prompt to instruct the model to first construct an egocentric map from the sampled frames, and then preserve the same coordinate-based egocentric reasoning pipeline used during training on MindCube-Tiny, as illustrated in Figure A5. This adaptation is important because VSI-Bench differs from our training setup in both the question format and the visual input distribution. Following the benchmark protocol, we report category-wise results together with the overall average score.

### S5.3 BLINK-MV

BLINK is a multiple-choice benchmark for visual perception that contains 14 tasks spanning low-level correspondence, geometric perception, and higher-level visual understanding. In this work, we use the multi-view reasoning subset, denoted as BLINK (MV), which evaluates whether a model can infer camera or scene changes from multiple images under partial and distributed visual evidence. As in the original benchmark, answer choices are discrete options and evaluation is based on exact-match accuracy after mapping the generated response to one of the candidate choices. We use the prompting template shown in Figure A6 and follow the standard BLINK evaluation protocol, reporting accuracy on this multi-view subset.

## S6 Additional Experimental Results

### S6.1 Additional Qualitative Results

We provide additional qualitative examples on all three benchmarks to illustrate how the proposed framework constructs intermediate spatial representations and performs grounded reasoning under diverse visual configurations. For MindCube, we show representative cases from both the *Among* and *Rotation* settings in Figures A7 and A8, highlighting how the model combines allocentric scene understanding with viewpoint-dependent egocentric grounding. Finally, Figure A9

**Prompt for BLINK (MV) evaluation**

**[System Prompt]**  
 You are a spatial reasoning assistant for multi-view 3D scenes. Given multiple images showing a 3D scene from different viewpoints, reason carefully about the spatial layout with constructing an allocentric cognitive map (CogMap) and answer the question. The CogMap is a structured grid where each object has an (x, y) grid coordinate in the allocentric (world-centered) frame, and each viewpoint has a position, facing direction, and associated image.

**[Answer Instruction]**  
 Output 3 blocks in order:

1. <CogMap>: Construct an allocentric cognitive map from multi-view images.
2. <think>: Complete a structured reasoning trajectory with 4 numbered steps:
  - (1) Selecting a view: Read the question, identify the spatial anchor, locate it in the CogMap, and determine which viewpoint is most relevant, e.g., We should consider the viewpoint from <Image X>.
  - (2) Egocentric layout: Convert the allocentric CogMap into an egocentric frame aligned with the selected viewpoint’s facing direction. Output as:
 

```
<EgoMap viewpoint="Image X">
- [object_name]: ego_direction="[direction]"
</EgoMap>
```

If rotation/movement is required, apply it and describe updated directions. For camera movement, compare the viewpoint positions across images in the CogMap to determine the direction of movement. Then, map the question’s spatial query to ego-direction labels, e.g., A.name=ego\_dir [Image X].
  - (3) Confirmation: Reference another image that visually confirms the answer, e.g., Confirming in <Image Y>: [object] is visible, consistent with the EgoMap.
  - (4) State the final answer: Therefore, the answer is “X. option\_text”.
3. <answer>: “X. option\_text”

**[CogMap Instruction]**  
 <cogmap\_gen\_instruction>

**[Egocentric Map Rules]**

- The viewpoint’s facing direction becomes “forward” in the EgoMap.
- Objects in the opposite direction become “back”.
- Apply 90° rotation for left/right accordingly.
- For diagonal facings, decompose and rotate consistently.
- Include only objects that are spatially relevant from the selected viewpoint.
- Rotation rules: Grid [0,0]=top-left, [9,9]=bottom-right.
  - facing=up → forward=[-y], right=[+x]
  - facing=right → forward=[+x], right=[+y]
  - facing=down → forward=[+y], right=[-x]
  - facing=left → forward=[-x], right=[-y]

**Fig. A6: Prompt template for BLINK (MV) evaluation.** We adapt a similar MindCube-Tiny prompt as shown in Figure A4 with additional guidance for camera movement questions, where the model compares viewpoint positions across images in the CogMap.

shows qualitative results on BLINK (MV), demonstrating that the proposed structured reasoning format also generalizes to multi-view perception problems beyond the original training benchmark. Overall, these examples qualitatively

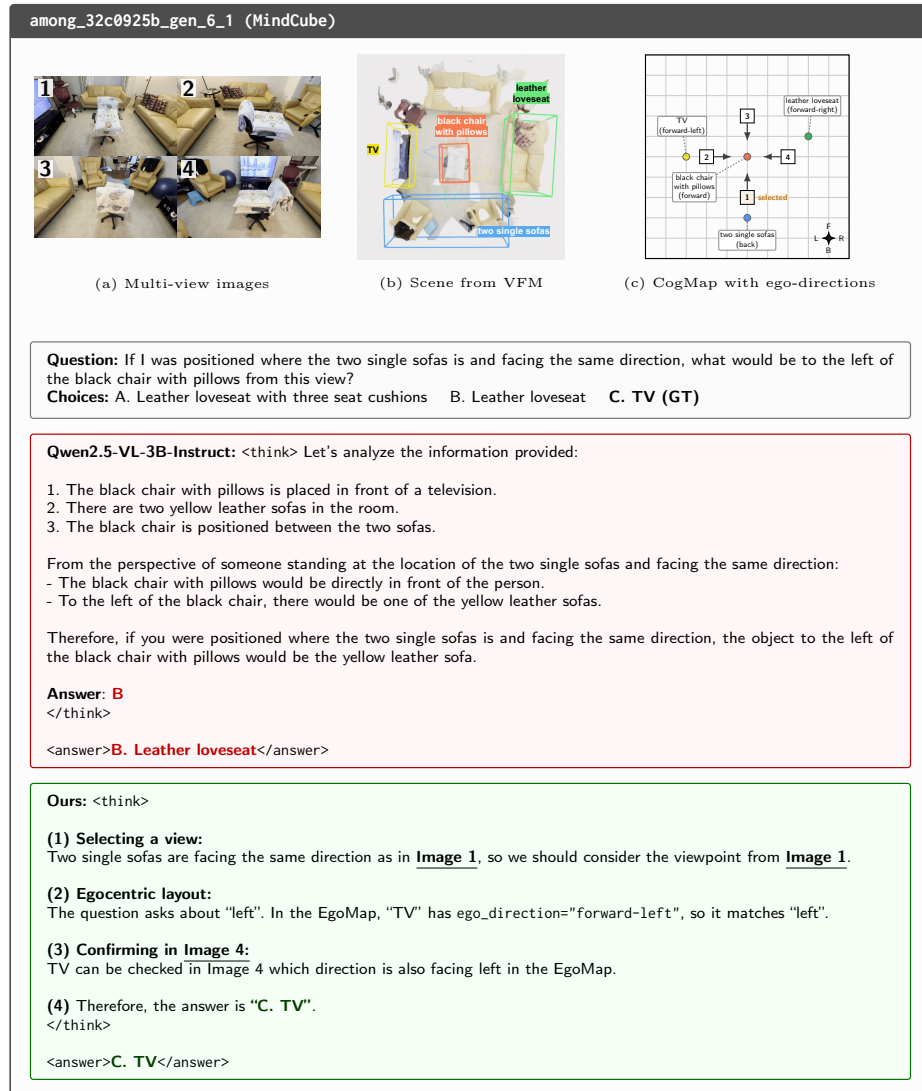


Fig. A7: Qualitative comparison on MindCube-Tiny (Among).

support that our method produces interpretable intermediate reasoning steps while remaining robust across different benchmark settings.

## S6.2 Training Plot and Reward Plot Analysis

We analyze training dynamics for both SFT and GRPO in our method. For SFT, we track the validation accuracy on MindCube-Tiny for up to 57 training

**Table A1: Analysis on effects of egocentric representation in inference using frozen MLLM (Qwen2.5-VL-3B-Instruct).**

Config.	Input		Output			MindCube-Tiny			
	CogMap	EgoMap	CogMap	EgoMap	Reasoning	Rotation	Among	Around	Overall
CogMap-FFR-Out	✗	✗	✓	✗	✓	25.00	39.67	58.40	41.33
CogMap-In-FFR-Out	✓	✗	✗	✗	✓	37.00	41.67	44.40	41.43
EgoMap-FFR-Out	✗	✗	✗	✓	✓	30.00	39.33	54.80	41.24
EgoMap-In-FFR-Out	✗	✓	✗	✗	✓	41.00	43.17	54.80	45.52
CogMap-EgoMap-FFR-Out	✗	✗	✓	✓	✓	33.50	45.50	54.80	45.43
CogMap-In-EgoMap-FFR-Out	✓	✗	✗	✓	✓	35.00	43.67	52.80	44.19
CogMap-EgoMap-In-FFR-Out	✓	✓	✗	✗	✓	55.50	51.17	52.00	<b>52.19</b>

steps. The left plot in Figure A10 compares three supervision settings: allocentric cognitive map (CogMap) only, trajectory only, and the full setting with both spatial maps and trajectory supervision. CogMap-only supervision converges to an average score of about 53, while trajectory-only supervision reaches about 58. The full setting achieves the highest plateau at about 62, which shows that spatial-map supervision and trajectory supervision are complementary. In particular, adding egocentric guidance with the trajectory of viewpoint selection leads to the greatest improvement in the final score.

For GRPO, the right plot in Figure A10 shows the mean reward over 200 training steps. We apply EMA smoothing to the step-wise reward because the raw reward is noisy across steps. We compare three reward settings: (i) the full reward, (ii) without  $\mathcal{R}_{\text{local}}$ , and (iii) without both  $\mathcal{R}_{\text{global}}$  and  $\mathcal{R}_{\text{local}}$ . Across all settings, the reward increases and then stabilizes, which indicates that GRPO optimization remains stable under each reward definition. The three curves follow a similar convergence pattern, but the full reward consistently stays above the ablated variants throughout training. This trend supports our reward design, as combining global and local rewards leads to the highest reward level under the same GRPO procedure.

### S6.3 Effects of Egocentric Grounding

We investigate the effect of egocentric grounding in an inference-only setting using a frozen MLLM (Qwen2.5-VL-3B-Instruct), without SFT or RL. We vary whether the model is (i) provided with an allocentric cognitive map (CogMap) or an egocentric map (EgoMap) as **inputs** (-In) and (ii) instructed to **generate** the maps as part of its **outputs** (-Out) with free-form reasoning (FFR).

As the results on MindCube-Tiny are summarized in Table A1, using only an allocentric cognitive map does not meaningfully change performance depending on whether it is provided as input or generated as output. Specifically, CogMap-FFR-Out and CogMap-In-FFR-Out achieve nearly identical overall scores with 41.33 and 41.43, respectively, and this supports that grounding an allocentric map alone is not the main driver of improvement in this zero-shot setting. In contrast, introducing an egocentric representation yields a clear gain. Providing an egocentric map as input (EgoMap-In-FFR-Out) improves the overall score to 45.52, and jointly providing both CogMap and EgoMap as inputs (CogMap-EgoMap-In-FFR-Out) further increases the overall score to 52.19. These

results suggest that egocentric grounding is important for spatial reasoning under viewpoint-dependent queries, and motivate our method to explicitly incorporate egocentric representations as grounding signals during training.

#### S6.4 Failure Cases

Despite the gains from map-grounded supervision and dense rewards, our method still has a clear limitation on scenes that violate the static-world assumption underlying our cognitive-map construction. In particular, the current framework assumes that the multi-view or multi-frame inputs can be integrated into a single coherent scene layout with stable object positions, so that camera motion can be inferred relative to a largely static environment. This assumption breaks when salient foreground objects move independently of the camera. In such cases, the model may incorrectly use a moving object as a spatial reference and attribute the apparent motion to ego-motion of the camera, leading to a wrong answer. For example, when a white car becomes closer across frames [53], the model predicts that the camera is moving forward, although the correct interpretation is that the object itself is moving and the camera is moving backward. This failure suggests that our current map-based reasoning is effective primarily for quasi-static scenes, but is less reliable when dynamic objects introduce ambiguity between object motion and camera motion. Addressing this limitation would require explicit motion disentanglement or dynamic scene representations beyond the static allocentric-egocentric pipeline considered in this work.

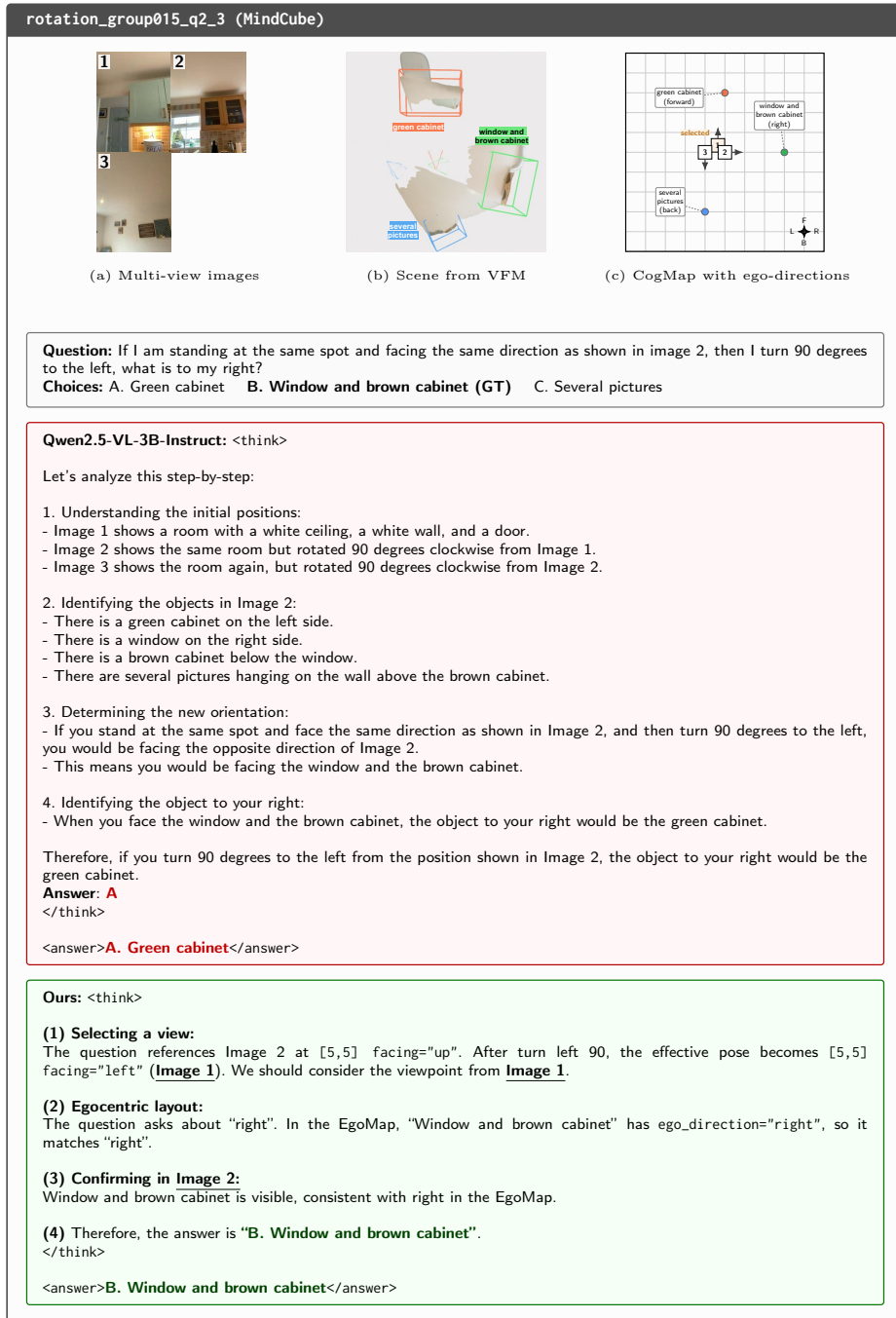




Fig. A8: Qualitative comparison on MindCube-Tiny (Rotation).

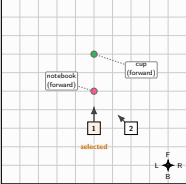
val\_Multi-view\_Reasoning\_67 (BLINK (MV))



(a) Multi-view images



(b) Scene from VFM



(c) CogMap with ego-directions

**Question:** The images are frames from a video. The first image is from the beginning of the video, and the second image is from the end. Is the camera moving left or right when shooting the video? Select between A and B.  
**Choices:** A. left B. right (GT)

**Qwen2.5-VL-3B-Instruct:** <think>  
 Let's break down this analysis:

**Image Description:**

- First Image:
  - There is an open laptop with its screen facing forward.
  - To the left of the laptop, there appears to be another monitor displaying some content.
- Second Image:
  - Similar setup but now we see more details such as someone sitting at the desk wearing jeans and sneakers.
  - Additional items visible include water bottles placed near the laptops.

**Analysis:**

- To determine if the camera moves left or right between these two images, let's consider several factors including perspective changes and relative positions observed across both shots.

**Perspective Changes:**

- In the initial frame, only one person seems present based solely on their partial view through the laptop screens.
- By the time it transitions into the next shot, additional elements come into focus. Notably showing part of another individual seated closer to us than before due likely to increased depth perception caused by movement towards them.

**Relative Position Shifts:**

- Initially focused mainly on objects behind individuals working; gradually shifts attention toward foreground figures, indicating a horizontal panning motion. The consistent framing implies a lateral shift along the X-axis, ruling out vertical rotation. The gradual repositioning from background to foreground suggests the camera is panning to the left.

**Answer: A**  
 </think>

<answer>A. left</answer>

**Ours:** <think>

**(1) Selecting a view:**  
 The question asks the movement direction from Image 1 to Image 2, so we should consider the viewpoint at Image 1.

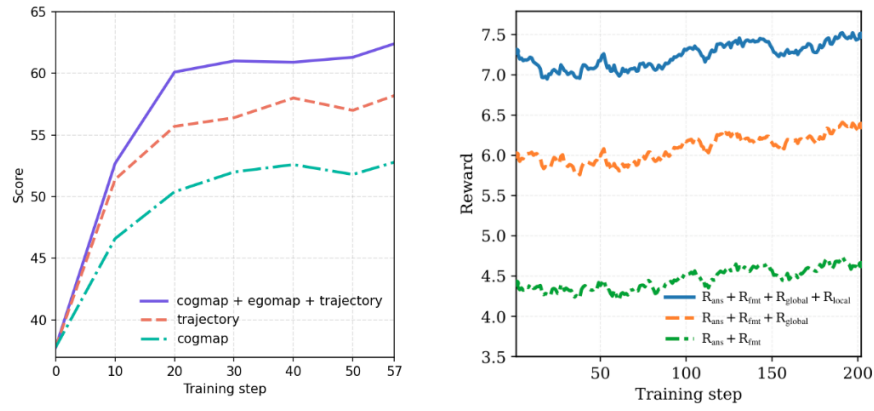
**(2) Egocentric layout:**  
 In the EgoMap, "notebook" has ego\_direction="forward". Also, the viewpoint at Image 2 is at the position (7,7). This corresponds to the "right".

**(3) Confirming in Image 2:**  
 "Notebook" appears on the left side of the cup, whereas it is on the "right" side in Image 1. This is consistent with a rightward camera movement.

**(4) Therefore, the answer is "B. right".**  
 </think>

<answer>B. right</answer>

**Fig. A9: Qualitative comparison on BLINK (MV).** The frozen Qwen2.5-VL-3B-Instruct fails to choose correct answer despite lengthy reasoning. In contrast, our method explicitly selects the appropriate reference view and resolves the camera motion with an ego-consistent reasoning trace, achieving the correct answer.



**Fig. A10: Training dynamics under different SFT and GRPO settings. Left:** During SFT, the average score increases and then saturates across settings, while the full setting reaches the highest plateau compared to variants only with trajectory supervision or spatial maps. **Right:** GRPO optimization proceeds stably under different reward settings, as all curves increase and then stabilize.