Spatial Reasoning in Large Vision-Language Models via Egocentric-to-Allocentric Transformation

Yoonji Kim[*1], **Jieun Kim**[*2], **Sung-Bae Cho**[1]
[1]Dept. of Computer Science, Yonsei University, Seoul, South Korea
[2]Dept. of Artificial Intelligence, Yonsei University, Seoul, South Korea
{yoonjikim,lilly9928,sbcho}@yonsei.ac.kr

Abstract

Spatial reasoning in large vision-language models involves inferring view-specific spatial relations and integrating them into a consistent scene-centered representation. However, existing approaches often learn view-dependent relations tied to individual camera frames, which limits their ability to form viewpoint-invariant spatial representations. Motivated by cognitive models of spatial perspective-taking, we propose **SCoRE** (Spatial Consistency Reasoning), a method that realizes multi-view spatial reasoning with egocentric-to-allocentric transformation, which first trains the model to infer egocentric object relations from single-view observations, and then applies reinforcement learning to encourage the model to estimate viewpoint orientation, transform egocentric relations into a shared scene-centered reference frame, and predict allocentric object relations. To enforce viewpoint-invariant reasoning, we optimize the model with a multi-view consistency reward that aligns predictions from different views in the common allocentric space. Experiments show that SCoRE improves spatial reasoning performance by +3.69%p on multi-view benchmarks and +4.72%p on general spatial reasoning benchmarks over the state-of-the-art methods, demonstrating the effectiveness of explicit reference-frame transformation for robust multi-view spatial understanding.

1 Introduction

Spatial reasoning is a fundamental capability for large vision-language models (LVLMs), enabling them to understand and interact with complex real-world environments (Stogiannidis et al., 2025; Jia et al., 2026; Li et al., 2025b). In such settings, LVLMs often observe scenes only partially and from varying viewpoints, requiring coherent spatial integration across views (Tao et al., 2026; Gholami et al., 2026). This capability is crucial for embodied navigation, robotics, and 3D scene understanding (Jeong et al., 2026; Feng et al., 2026); however, despite recent progress, LVLMs still struggle to reason consistently across partial observations and viewpoints (Yin et al., 2025; Xue et al., 2025). Existing approaches primarily rely on direct supervision of final answers, including supervised fine-tuning (Cai et al., 2026) and reinforcement learning (Li et al., 2026; Wu et al., 2025; Pan and Liu, 2026) with task-specific rewards. However, such supervision mainly encourages models to produce correct final outputs without explicitly constraining how spatial relations should be transformed across viewpoints(Kim et al., 2026a). As illustrated in Figure 1, an LVLM may answer spatial questions correctly from one view but produce inconsistent predictions

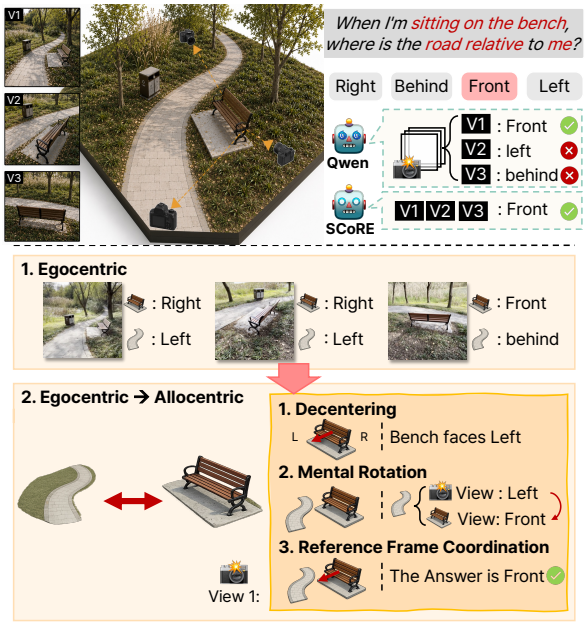


Figure 1: **Egocentric-to-Allocentric Spatial Reasoning.** Existing LVLMs rely on egocentric cues, causing inconsistent predictions across views of the same scene. SCoRE instead transforms egocentric observations into a unified allocentric representation.

[*]Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 38100–38118
October 24-29, 2026 ©2026 Association for Computational Linguistics

from other views of the same scene. This suggests that LVLMs often rely on viewpoint-specific egocentric observations rather than a shared spatial representation, leading to viewpoint-dependent shortcuts.

Humans naturally address this challenge by systematically transforming egocentric observations into a consistent allocentric representation (Siegel and White, 1975; Klatzky; O'Keefe and Nadel, 1978). In contrast, current LVLMs do not explicitly learn such transformation processes, leaving a critical gap between spatial perception from individual views and viewpoint-invariant spatial reasoning.

To address this gap, we propose **SCoRE** (Spatial Consistency Reasoning), a model that explicitly learns egocentric-to-allocentric transformation as a structured reasoning process. As shown in Figure 1, SCoRE first identifies egocentric spatial relations from each partial observation and then transforms them into a shared allocentric frame through decentering, mental rotation, and reference-frame coordination. We train SCoRE with a multi-stage strategy: supervised fine-tuning for egocentric spatial perception, transformation fine-tuning for structured reasoning traces, and Group Relative Policy Optimization (GRPO) with a composite reward. In particular, multi-view consistency serves as the primary supervision signal, encouraging viewpoint-specific observations of the same scene to converge to a shared allocentric relation. This discourages viewpoint-dependent shortcuts and promotes robust spatial reasoning across views.

Experimental results show SCoRE outperforms prior SOTA by +3.69%p on multi-view and +4.72%p on general spatial reasoning benchmarks, demonstrating that explicit view transformation enhances spatial reasoning ability.

- **Egocentric-to-Allocentric Reasoning**: We propose SCoRE, a method that explicitly learns view transformation for spatial reasoning.
- **Viewpoint-Invariant Training**: We propose a multi-stage training strategy that combines structured transformation learning with consistency-based GRPO.
- **Generalizable Spatial Reasoning**: We show that multi-view consistency learning improves viewpoint-invariant spatial understanding beyond multi-view settings.

2 Related Work

Spatial Representations in Human Cognition. Humans achieve robust spatial understanding through two complementary representations: egocentric representations centered on the observer's viewpoint, and allocentric representations that encode spatial relationships independently of the observer (Klatzky; Burgess, 2006). Spatial information is first perceived egocentrically, relative to one's own position and orientation (Wang and Spelke, 2000, 2002), and is then transformed into allocentric frames to remain stable across viewpoint changes (Byrne et al., 2007). These transformations support cognitive maps, which encode stable relationships among places and objects for navigation and wayfinding (O'Keefe and Nadel, 1978; Moser et al., 2008). In contrast, current LVLMs often remain tied to viewpoint-specific visual patterns, without explicitly modeling how spatial relations change with viewpoint. This motivates our approach to model egocentric-to-allocentric transformations as a structured spatial reasoning process.

Spatial Reasoning in LVLMs. Spatial reasoning in LVLMs has gained increasing attention for robust real-world reasoning, with existing methods primarily relying on prompting, supervised fine-tuning (SFT), and reinforcement learning (RL). Prompting-based methods elicit spatial reasoning without additional training (Yang et al., 2025), but they remain sensitive to prompt formulation and vary significantly across model architectures. SFT-based methods have recently explored directions such as synthetic spatial supervision (Boyuan Chen, 2024) and large-scale spatial instruction tuning (Cai et al., 2026). However, they remain constrained by supervision quality and coverage, relying on pattern matching rather than deeper spatial understanding. RL-based methods instead optimize spatial reasoning through reward mechanisms, with recent works introducing verifiable spatial rewards (Pan and Liu, 2026; Wang and Ling, 2025). However, all three paradigms largely treat spatial reasoning as an implicit capability, risking dataset-specific shortcuts rather than genuine spatial understanding, where spatial relationships should remain consistent regardless of how a scene is observed. To address this limitation, SCoRE trains models to reason under spatial consistency constraints.

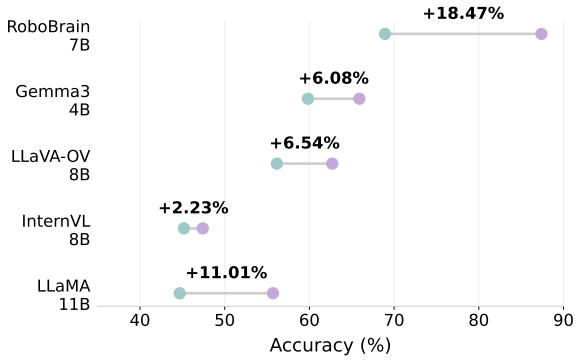


Figure 2: **Spatial Perspective Transformation with CoT.** Accuracy of five LVLMs on the multi-view dataset without prompting (● *Origin*) and with egocentric-to-allocentric CoT prompting (● *Ego-to-Allo CoT*).

3 Methodology

Section 3.1 presents the formulation, and Section 3.2 introduces our two-stage training strategy.

3.1 Spatial Reasoning Formulation

Humans reason about space by transforming egocentric observations into allocentric representations, enabling stable scene-level understanding across viewing directions. We formulate spatial reasoning in LVLMs through an egocentric-to-allocentric transformation.

As shown in Figure 2, prompting LVLMs to reason through this transformation improves spatial reasoning performance. However, these gains are inconsistent and sensitive to prompt formulations (Appendix B.2).

We therefore incorporate the egocentric-to-allocentric transformation into training and formalize it with four reasoning components:

$$R = \{r^{\text{ego}}, r^{\text{head}}, r^{\text{trans}}, r^{\text{allo}}\}, \tag{1}$$

where r^{ego} describes the egocentric observation, r^{head} specifies the facing direction of the subject, r^{trans} performs the allocentric spatial transformation, and r^{allo} derives the final conclusion.

Based on this structure, we construct training reasoning traces consisting of view-wise reasoning steps followed by a final allocentric conclusion. Additional dataset details are provided in Appendix A.2.

3.2 Spatial Consistency Reasoning

Stage 1: Egocentric Spatial Grounding. The first stage trains the model to acquire basic camera-centric spatial perception through supervised fine-tuning. This reflects the intuition that robust spatial reasoning first requires recognizing spatial evidence from the current viewpoint before transforming it into an allocentric representation. Given a single-view perception dataset

$$\mathcal{D}_{\text{perc}} = \{(I_i, Q_i, Y_i)\}_{i=1}^{N_p}, \tag{2}$$

where I_i denotes a single input image, Q_i is a multiple-choice spatial question, and Y_i is a directional relation label such as *front*, *behind*, *left*, or *right*, we train the model to directly predict the correct relation label.

This stage uses only single-view inputs and does not require the model to generate explicit reasoning traces. Instead, it provides a perceptual foundation for the subsequent transformation stage by grounding the model in elementary spatial relation recognition.

Stage 2: Perspective-Taking Reasoning. The second stage learns egocentric-to-allocentric transformation for spatial reasoning. Starting from the perception checkpoint obtained in Stage 1, we first employ SFT for the egocentric-to-allocentric transformation as a cold start for the structured reasoning traces defined in Section 3.1.

Given the transformation SFT dataset

$$\mathcal{D}_{\text{trans}} = \{(\mathcal{I}_i, Q_i, R_i, Y_i)\}_{i=1}^{N_t}, \tag{3}$$

where $\mathcal{I}_i = \{I_i^{(v)}\}_{v=1}^{V}$ denotes multiple views of the same scene, Q_i is the spatial reasoning question, R_i is the view-wise egocentric-to-allocentric reasoning trace, and Y_i is the final viewpoint-invariant answer. We train the model to autoregressively generate R_i followed by Y_i.

After transformation SFT, we further optimize the model using GRPO with cross-view consistency rewards. While SFT provides a cold start for the structured transformation format, egocentric-to-allocentric transformation does not correspond to a single unique reasoning trajectory. Thus, supervised learning alone may encourage the model to imitate dataset-specific trace patterns or shortcut to the final answer without reliably performing view-wise transformation.

GRPO addresses this limitation by sampling multiple reasoning trajectories and assigning higher rewards to trajectories whose view-wise transformed relations are consistent with the shared allocentric ground truth. Unlike final-answer-only supervision,

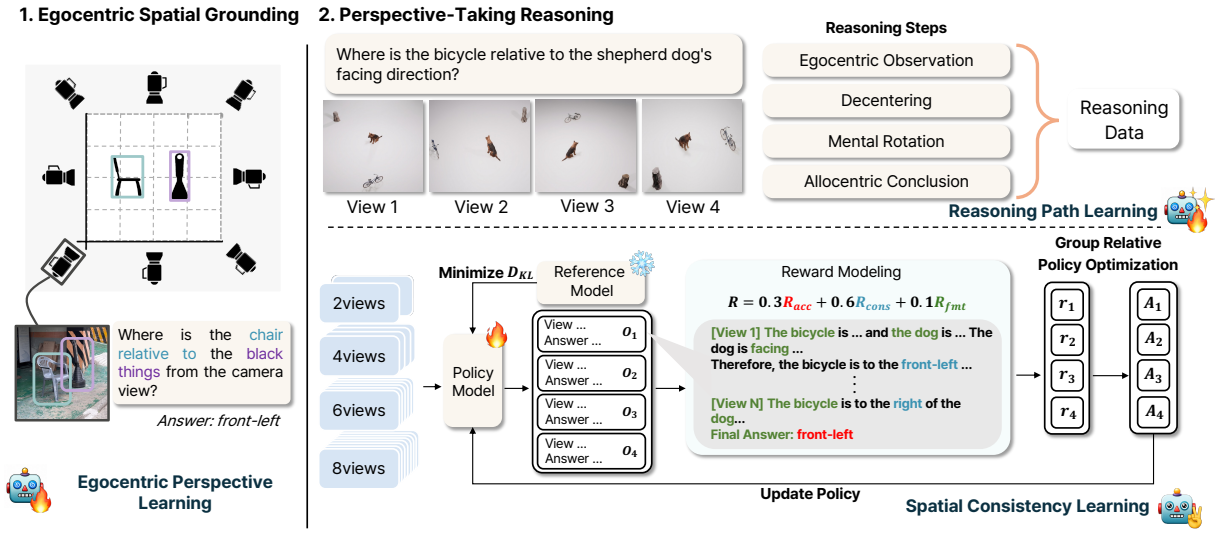


Figure 3: **Overview of SCoRE.** Stage 1 builds camera-centric spatial perception from single-view observations. Stage 2 teaches egocentric-to-allocentric transformation through transformation SFT, then further optimizes the policy with GRPO using cross-view consistency rewards for viewpoint-invariant spatial reasoning.

this reward discourages shortcut solutions and promotes stable reasoning across diverse camera-view subsets.

Each training sample is represented as $(\mathcal{I}_s, Q_s, Y_s, \Theta_s)$, where $\mathcal{I}_s$ contains multiple observations associated with a spatial reasoning example, Q_s is the question, Y_s is the ground-truth allocentric answer, and Θ_s denotes the corresponding camera headings.

Given an input sample, the model generates G sampled reasoning trajectories:

$$O_i \sim \pi_\phi(\cdot \mid \mathcal{I}_s, Q_s, \Theta_s), \quad i = 1, \ldots, G, \tag{4}$$

where we use $G = 4$ sampled outputs during training.

For each sampled trajectory, we extract the predicted final answer:

$$\hat{Y}_i = \mathrm{Ext}(O_i). \tag{5}$$

The reward is defined as a weighted combination of correctness, cross-view consistency, and format quality. As enforcing viewpoint-invariant reasoning through cross-view consistency is the central goal of our training objective, we assign it the largest weight and formulate the reward as:

$$R_i = 0.3R_{\text{acc}} + 0.6R_{\text{cons}} + 0.1R_{\text{fmt}}. \tag{6}$$

The correctness reward evaluates whether the predicted final answer matches the ground-truth label:

$$R_{\text{acc}} = \begin{cases} 1, & \hat{Y}_i = Y_s \\ -1, & \hat{Y}_i \neq Y_s \end{cases} \tag{7}$$

The consistency reward measures whether view-wise transformed conclusions remain consistent and agree with the target allocentric relation.

The set $\mathcal{V}_i$ is constructed by parsing the generated reasoning trajectory O_i view by view. We first split O_i into blocks of the form `View N:`. For each block, we only consider the text following `Therefore`, preventing egocentric relations mentioned in earlier observation or heading-identification steps from being confused with the final transformed allocentric conclusion. A valid spatial relation is then extracted from each view block, and the resulting view-wise conclusions form $\mathcal{V}_i$.

The consistency reward is defined as:

$$R_{\text{cons}} = \begin{cases} R_{\text{match}} - 0.3R_{\text{uniq}}, & |\mathcal{V}_i| > 0, \\ -0.5, & |\mathcal{V}_i| = 0, \end{cases} \tag{8}$$

where

$$R_{\text{match}} = \frac{\sum_{y \in \mathcal{V}_i} \mathbb{I}[y = Y_s]}{|\mathcal{V}_i|}, \tag{9}$$

and

$$R_{\text{uniq}} = \frac{|\mathrm{Unique}(\mathcal{V}_i)| - 1}{|\mathcal{V}_i|}. \tag{10}$$

The format reward verifies whether the generated reasoning follows the required output structure or not. It checks for the presence of four components, assigning 0.25 points to each: view decomposition, image-location description, heading identification, and final-answer reporting.

We optimize the policy using GRPO. For a group of G sampled outputs, the normalized group-relative advantage is:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G) + \epsilon}. \tag{11}$$

We optimize the policy with

$$\mathcal{J}_{\text{GRPO}}(\phi) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1\pm\epsilon)\hat{A}_i\Big) - \beta D_{\text{KL}}(\pi_\phi \| \pi_{\text{ref}})\right] \tag{12}$$

where

$$\rho_i = \frac{\pi_\phi(O_i \mid \mathcal{I}_s, Q_s, \Theta_s)}{\pi_{\phi_{\text{old}}}(O_i \mid \mathcal{I}_s, Q_s, \Theta_s)} \tag{13}$$

We use clipping threshold $\epsilon = 0.2$ and KL coefficient $\beta = 0.04$ to preserve SFT-initialized reasoning behavior while preventing policy collapse.

4 Experiments

4.1 Experimental Setup

Training Settings. We use Qwen2.5-VL-7B as the base model and build on SCOPE (Kim et al., 2026b), which provides scene-level spatial metadata and camera pose information. SCoRE is trained through a multi-stage pipeline. Detailed settings are provided in Appendix A.

Baselines. Our baselines include SoTA GRPO-based spatial reasoning models and base LVLM backbones (Bai et al., 2025). All models are evaluated using the system prompts and settings specified in their corresponding papers. VILASR (Wu et al., 2025) uses drawing-based intermediate representations. SpatialReasoner (Ma et al., 2025) performs explicit 3D coordinate reasoning. Spatial-Thinker (Batra et al., 2025) uses scene graph representations. Spatial-SSRL (Liu et al., 2026) learns spatial representations through self-supervised spatial reconstruction.

Benchmarks. We evaluate on six benchmarks covering in-domain allocentric reasoning and out-of-domain general spatial reasoning.

In-domain. ViewSpatial-Bench (Li et al., 2025b) evaluates multi-perspective spatial localization, MMSI-Bench (Yang et al., 2026) focuses on multi-image spatial intelligence, and MindCube (Yin et al., 2026) assesses spatial mental modeling from limited views.

Out-of-domain. What'sUp (Kamath et al., 2023) evaluates single-image 2D spatial relations, while SPAR-Bench (Zhang et al., 2025) and BLINK (Fu et al., 2024) cover broader spatial perception and reasoning tasks. We use these benchmarks to evaluate the generalizability of SCoRE beyond allocentric settings.

4.2 Main Results

Multi-View Spatial Reasoning. As shown in Table 1, SCoRE achieves the highest average accuracy on both MindCube (39.28%) and MMSI-Bench (30.50%), achieving the best performance among existing methods. The improvements are particularly evident on viewpoint-sensitive subtasks. On MindCube, SCoRE achieves the best performance on *Among* and substantially improves performance on *Around*. Both tasks require maintaining spatial consistency across changing views, aligning well with SCoRE's egocentric-to-allocentric transformation. On MMSI-Bench, SCoRE shows especially strong gains on *MSR*, suggesting improved consistency in multi-step spatial reasoning. Overall, the results indicate that SCoRE's consistency-oriented training particularly benefits tasks requiring stable allocentric representations and multi-step egocentric-to-allocentric transformations.

Egocentric - Allocentric Reasoning. As shown in Figure 4, SCoRE achieves the highest overall accuracy (42.65%) among SoTA models on a 1,000-sample subset of ViewSpatial-Bench, outperforming Spatial-SSRL by 4.60 percentage points. Although SCoRE is trained with both egocentric and allocentric supervision, its gains are particularly pronounced in allocentric reasoning, where it reaches 46.31% and surpasses the strongest prior model by 3.88 percentage points. Compared to the Qwen baseline, SCoRE improves allocentric performance by 11.14 points, compared with a 2.83-point improvement in egocentric performance. Since allocentric reasoning requires maintaining spatial relationships across viewpoint changes and reasoning beyond the observer's own perspective, these results suggest that SCoRE's training strategy is particularly effective for challenging viewpoint-dependent spatial reasoning.

General Spatial Reasoning. Table 2 evaluates whether consistency-oriented training generalizes beyond the spatial capabilities explicitly targeted during training. SCoRE achieves the best perfor-

Table 1: **Main results on multi-view spatial reasoning benchmarks.** QA accuracy (%) on MindCube and MMSI-Bench. Avg. denotes the average accuracy within each benchmark, and Δ denotes the absolute improvement over the base model.

Model	MindCube				MMSI-Bench					Avg.	Δ
	Rotation	Among	Around	Avg.	Positional	Attribute	Motion	MSR	Avg.		
Qwen2.5-VL-7B	31.00	30.17	27.25	29.47	28.87	25.53	20.61	19.70	25.40	27.44	–
Baselines											
VILASR-7B	33.00	28.83	25.00	28.94	33.36	28.60	18.62	24.75	28.80	28.87	+1.43
SpatialReasoner	32.50	38.67	29.25	34.50	20.87	27.82	19.33	24.75	22.20	28.35	+0.91
SpatialThinker	35.00	30.50	25.25	30.25	28.37	27.75	26.57	26.26	28.14	29.20	+1.76
Spatial-SSRL	39.50	29.00	25.75	31.42	28.22	21.74	25.30	27.78	26.43	28.93	+1.49
Ours											
SCoRE(w/o RL)	34.00	39.33	42.50	38.61	28.62	23.91	19.35	24.75	25.80	32.21	+4.77
SCoRE (full)	34.50	40.83	42.50	39.28	29.06	27.65	26.64	39.39	30.50	34.89	**+7.45**

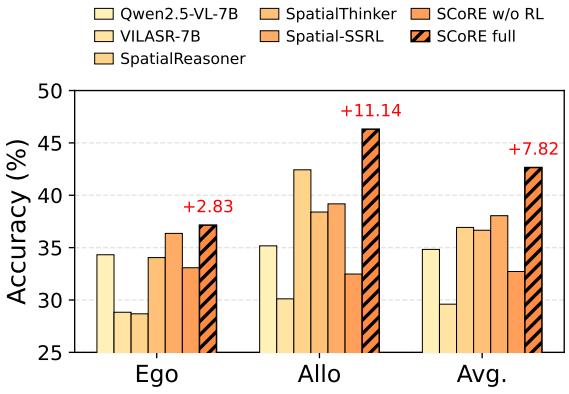


Figure 4: **Evaluation on ViewSpatial-Bench.** QA accuracy (%) of different models on Ego, Allo, and Avg. settings.

Table 2: **Evaluation on general spatial reasoning benchmarks.** QA accuracy (%). Avg. denotes the average accuracy across benchmarks, and Δ denotes the absolute improvement over the base model.

Model	What's Up	SPAR-Bench	BLINK	Avg.	Δ
Qwen2.5-VL-7B	90.10	33.85	69.00	64.32	–
Baselines					
SpatialReasoner	84.20	33.21	49.34	55.58	-8.74
SpatialThinker	92.30	34.04	65.20	63.85	-0.47
Ours					
SCoRE (w/o RL)	97.60	34.10	71.75	67.82	+3.50
SCoRE (full)	98.00	35.47	72.25	68.57	**+4.25**

mance across all three benchmarks, improving the average accuracy from 64.32% to 68.57% over the base model. Notably, on What'sUp, which evaluates relative spatial understanding from single-view images, SCoRE reaches 98.00%, suggesting that multi-view consistency training also benefits fine-grained single-view spatial reasoning. SCoRE also improves SPAR-Bench from 33.85% to 35.47% and BLINK(val-spatial) from 69.00% to 72.25%.Overall, these results suggest that SCoRE's egocentric-to-allocentric training improves general spatial representations, leading to gains beyond the multi-view setting used for training.

Category-Level Analysis. Figure 5 compares model performance across four spatial reasoning categories. SCoRE shows clear gains in key subcategories, especially those involving agent–object relations and viewpoint transitions. In Relation Pattern, SCoRE substantially improves *A-O* reasoning (44.73% vs. 23.65% and 23.91%), suggesting that it better captures spatial relations between an agent and surrounding objects. In Viewpoint Dynamics, SCoRE also achieves a large gain on *Sequence* (50.17% vs. 24.42% and 24.75%), indicating stronger ability to track how the perceived scene changes across successive viewpoints. These improvements are notable because both settings require reasoning beyond static object recognition: the model must infer relations from an embodied reference frame and update them as the viewpoint changes. Overall, the results suggest that SCoRE is particularly effective for spatial reasoning tasks that require relational understanding and dynamic viewpoint tracking.

4.3 Ablation Studies

Component Ablation. Table 3 analyzes Perception SFT (Perc. SFT), Transformation SFT (Trans. SFT), and GRPO-based consistency refinement in SCoRE. The partial-training results

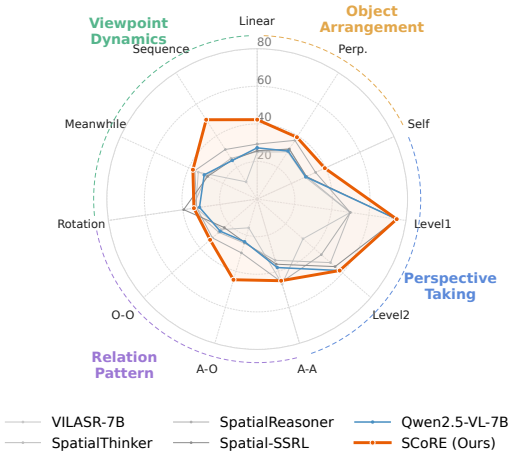


Figure 5: **Category-Level Analysis on MindCube.** Fine-grained comparison across all 11 MindCube subcategories between SCoRE, the base model, and existing baseline methods.

show that no single stage recovers the full performance: RL alone achieves the lowest average accuracy (28.24%), whereas Trans. SFT performs best among the single-stage variants (29.47%). Removing GRPO from the full pipeline causes the largest average drop, from 34.89% to 32.21%, suggesting that consistency refinement benefits from structured SFT initialization. Removing Trans. SFT also has a larger impact than removing Perc. SFT, supporting the importance of explicit egocentric-to-allocentric transformation in the staged pipeline. The full SCoRE achieves the highest average accuracy (34.89%).

Reward Ablation. Table 4 shows the results of an ablation study on the GRPO reward, which combines answer correctness, cross-view consistency, and output format rewards ($0.3\,R_{\text{acc}} + 0.6\,R_{\text{cons}} + 0.1\,R_{\text{fmt}}$); to isolate the contribution of each component, we remove the correctness reward (w/o Corr) and the consistency reward (w/o Cons) separately, in addition to the previously reported w/o RL setting. The results show that removing R_{cons} leads to the largest performance drop (-1.61%p on MindCube, -4.60%p on MMSI), while removing R_{acc} also degrades performance but to a lesser extent (-0.45%p on MindCube, -3.60%p on MMSI). This asymmetry suggests that correctness alone is insufficient to guide the model toward stable multi-view reasoning: without an explicit signal encouraging agreement across views, the model is more likely to arrive at correct answers through inconsistent or view-dependent reasoning paths that generalize poorly, particularly on MMSI-Bench, which involves more diverse camera configurations than MindCube.

Table 3: **Component ablation study.** QA accuracy (%) on Qwen2.5-VL-7B.

Training setup	MindCube	MMSI	Avg.
Partial training			
Perc. SFT only	34.03	22.90	28.47
RL only	30.25	26.23	28.24
Trans. SFT only	31.09	27.84	29.47
Ablation			
w/o Perc. SFT	36.87	31.32	34.10
w/o Trans. SFT	35.76	30.49	33.13
w/o RL	38.61	25.80	32.21
SCoRE (full)	39.28	30.50	34.89

Table 4: **Ablation of GRPO reward components.** Removing the consistency reward causes the largest accuracy drop, confirming it as the primary driver of SCoRE's gains. QA accuracy (%) on MindCube and MMSI-Bench.

Setting	Corr.	Cons.	Fmt.	MindCube	MMSI
w/o Cons	✓		✓	37.67	25.90
w/o Corr		✓	✓	38.83	26.90
Full	✓	✓	✓	**39.28**	**30.50**

Reward Dynamics. Table 5 compares reward dynamics during GRPO training across different SFT initialization variants. We compute early and late rewards as the mean rewards over the first 200 and last 200 optimization steps, respectively, among 1,032 total training steps. Models initialized with Transformation SFT yield substantially higher early rewards (+0.52 and +0.60) compared to those without (-0.47 and -0.58), confirming that transformational perspective training provides an initialization for RL optimization. Although the reward improvement (Δ) of our full model (Perception+Transformation SFT $\rightarrow$ GRPO) is modest (+0.03), this is consistent with the observation that a well-initialized model requires less correction during RL training. Crucially, this reward improvement translates into consistent accuracy gains on downstream benchmarks (Table 1), validating the effectiveness of GRPO as the final refinement stage in SCoRE.

Table 5: **Reward Dynamics during GRPO Training.** Early and late reward comparison across initialization variants.

SFT Initialization		Reward		Δ
Perception	**Transformation**	**Early**	**Late**	
		-0.47	-0.46	+0.01
✓		-0.58	-0.35	+0.24
	✓	+0.52	+0.53	+0.01
✓	✓	+0.60	+0.62	+0.03

Table 6: **Relational consistency evaluation on MindCube-Hard.** All metrics are reported in %. Values in parentheses denote absolute improvements over the base model.

Model	**Relation Accuracy**	**Pair Coverage**	**Global Consistency**
Qwen2.5-VL-7B	48.50	45.93	23.00
Baselines			
SpatialReasoner	36.74 (-11.76)	40.33 (-5.60)	19.00 (-4.00)
SpatialThinker	50.67 (+2.17)	49.40 (+3.47)	27.00 (+4.00)
Spatial-SSRL	51.23 (+2.73)	46.37 (+0.44)	32.00 (+9.00)
SCoRE	52.47 (+3.97)	50.33 (+4.40)	34.00 (+11.00)

5 Consistency and Robustness Analysis

Relational Consistency Evaluation. We evaluate relational consistency by converting MindCube into a relational QA setting, where predictions are aggregated into pairwise spatial constraints and compared with the ground-truth object configuration. We use three metrics: *Relation Accuracy*, which measures the correctness of inferred pairwise relations; *Pair Coverage*, which measures the proportion of object pairs with valid inferred relations; and *Global Consistency*, which measures the proportion of samples where all evaluated relations are correct. Details are provided in Appendix B.3. As shown in Table 6, SCoRE achieves the best performance across all metrics. Compared with Qwen2.5-VL-7B, SCoRE improves Relation Accuracy by +3.97%p, Pair Coverage by +4.40%p, and Global Consistency by +11.00%p. These results show that SCoRE infers more accurate and broader pairwise relations while better preserving scene-level consistency. The large gain in Global Consistency suggests that the proposed method improves structured spatial reasoning, where multiple local relation predictions must align with a coherent object layout.

Table 7: **Step-wise reasoning trace analysis.** Accuracy of each intermediate reasoning step defined in Eq. (1).

Reasoning Step	**Accuracy (%)**
Egocentric relation (r^{ego})	84.5
Heading (r^{head})	81.2
Transformation (r^{trans})	98.9
Allocentric relation (r^{allo})	68.1

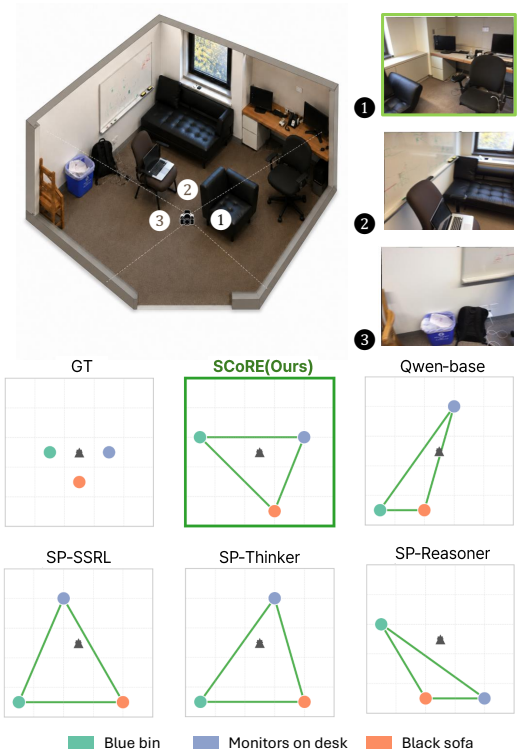


Figure 6: **Visualization of relational consistency.** SCoRE produces a coherent object layout consistent with the ground truth, unlike baseline models.

Step-wise Reasoning Trace Analysis. We further analyze the reasoning traces generated during GRPO training by evaluating each intermediate step of Eq. (1) separately: egocentric relation recognition, heading estimation, and the coordinate transformation itself. As shown in Table **??**, the transformation consistency reaches 98.9%, indicating that the model reliably composes the predicted egocentric relation and heading into the correct allocentric relation, rather than the observed gains arising from spurious correlations in the final answer alone.

Case Study. Figure 6 visualizes a qualitative comparison of spatial relationship reconstruction across different methods. Given multiple egocentric observations, SCoRE accurately recovers the



Question: When I am using the handwashing area, where is the toilet area located in relation to me?

A: To my right, B: Directly in front of me, C: Behind me, **D: To my left**

SCoRE:
Image 1: There is a towel hanging on the left side of the image. The sink is centrally located in the image. **When heading toward the handwashing area, the towel is on the front-left side.**
Image 2: There is a towel hanging on the right side of the image. The toilet is centrally located in the image. **The toilet is on the left side of the towel.**
Now, in Image 1, when facing the handwashing area, the towel is located to the front-left of the handwashing area. In Image 2, the toilet is to the left of the towel. Therefore, the toilet area is **to my left.**

Qwen:
The image shows a bathroom with a sink on the left side and a toilet on the right side. The sink has a countertop with a mirror above it, and there is a towel hanging on the left side of the sink. The toilet is positioned against the wall, and there is a towel rack with a towel hanging on the right side of the image.
From the camera's viewpoint, **if you were standing at the sink, the toilet would be to your right.** Therefore, the toilet area is located to my **right.**

Figure 7: **Visualization of Success Case.** Blue highlights indicate correct reasoning steps, and red highlights indicate incorrect reasoning.

globally consistent allocentric layout, achieving the highest similarity to the ground truth. In contrast, baseline methods often distort the global geometry or collapse object relations into incorrect triangular layouts. As shown in Figure 7, we present a representative example from MindCube to illustrate the reasoning behavior. SCoRE reasons from the viewpoint of the target person inside the scene, using the person's facing direction as the reference frame rather than treating the image as a fixed observer view. It composes spatial relations across views and correctly concludes that the toilet is to the user's left. In contrast, Qwen recognizes several local visual cues, such as the relative positions of the sink, towel, and toilet, but fails to consistently transform them into the user's facing direction, leading to an incorrect right-side conclusion.

Robustness to Spatial Reference Ambiguity. Since explicit viewpoint and orientation information may not always be available in real-world settings, we evaluate SCoRE under three input conditions on MindCube. In the *explicit* setting, each view is annotated with its camera viewpoint and heading, following the original MindCube protocol. In the *absent* setting, these annotations are removed entirely, requiring the model to infer viewpoint and orientation cues directly from the visual content. In the *noisy* setting, the viewpoint/orientation labels are deliberately corrupted (e.g., swapped or randomly perturbed) to simulate unreliable or mismatched metadata that may arise in practice. Table 8 shows that SCoRE achieves its best performance (39.67%) in the absent setting and shows only modest degradation both with explicit information (35.33%) and under noisy input (34.33%), consistently outperforming all baselines across all three settings. In contrast, baselines such as Spatial Reasoner exhibit noticeably larger performance gaps across the three conditions, suggesting greater sensitivity to the presence or quality of viewpoint information. These results indicate that explicit viewpoint/orientation descriptions are not a prerequisite for SCoRE's spatial reasoning and may even introduce unnecessary linguistic overhead.

Table 8: **Robustness to viewpoint/orientation (V&O) uncertainty.** SCoRE consistently outperforms all baselines across explicit, absent, and noisy V&O settings, showing that explicit viewpoint descriptions are not a prerequisite for its spatial reasoning.

Model	w/o V&O	w/ V&O	w/ Noise
Qwen2.5-VL-7B	24.00	25.00	25.33
Spatial Reasoner	32.33	29.33	30.67
Spatial Thinker	26.33	26.00	28.00
VILASR-7B	22.00	22.00	20.00
Spatial-SSRL	29.00	29.67	30.00
SCoRE	**39.67**	**35.33**	**34.33**

6 Concluding Remarks

This paper addresses the challenge of multi-view spatial reasoning with SCoRE, a model that learns consistent allocentric representations through explicit egocentric-to-allocentric transformation. By facilitating viewpoint-invariant reasoning across observations, SCoRE aligns spatial relations within a shared scene-centered reference frame, reducing reliance on view-dependent representations. Experimental results show consistent improvements over prior SoTA methods across multi-view benchmarks, general spatial reasoning tasks, and even single-view spatial reasoning. These findings suggest that enforcing multi-view spatial consistency helps models learn more generalizable spatial representations, highlighting a promising direction for robust spatial reasoning in vision-language models.

7 Limitations

SCoRE currently focuses on allocentric representations, relying on available supervision for egocentric-to-allocentric transformation. While SCoRE demonstrates strong performance on existing benchmarks, the diversity of training data remains limited. Expanding supervision to cover more diverse real-world viewpoints and scene configurations would further strengthen egocentric-to-allocentric transformation and broaden its applicability to more varied spatial reasoning scenarios.

8 Ethical Considerations

This work focuses on spatial reasoning for vision-language models and does not involve any personally identifiable information or human subjects. All training data is either synthetically generated via Blender-based rendering or derived from publicly available benchmarks under permissive licenses. We do not foresee direct ethical risks associated with the proposed method; however, as with any vision-language model, we encourage careful evaluation before deployment in safety-critical real-world applications.

Acknowledgments

This work was supported by IITP grant funded by the Korea government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University); No. RS-2022-II220113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming Yang, Zhiyuan Li, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Hunar Batra and 1 others. 2025. Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. In *NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI*.

Sean Kirmani Brain Ichter Dorsa Sadigh Leonidas Guibas Fei Xia Boyuan Chen, Zhuo Xu. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Neil Burgess. 2006. Spatial memory: how egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10(12):551–557.

Patrick Byrne, Suzanna Becker, and Neil Burgess. 2007. Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, 114(2):340–375.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, and 10 others. 2026. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

ZhiYuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shi Shubin, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, Rushuai Yang, Ruichuan An, Leqi Zheng, Weijie Wang, Shuang Chen, Sicheng Xu, Yaobo Liang, Jiaolong Yang, and Baining Guo. 2026. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. In *The Fourteenth International Conference on Learning Representations*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv:2503.19786*.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. 2026. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *The Fourteenth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv:2407.21783*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Suchae Jeong, Jaehwi Song, Haeone Lee, Hanna Kim, Jian Kim, Dongjun Lee, Dong Kyu Shin, Changyeon Kim, Dongyoon Hahm, Woogyeol Jin, and 1 others. 2026. Learning multi-view spatial reasoning from cross-view relations. *arXiv preprint arXiv:2603.27967*.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. 2026.

Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's up? adapting vision-language models for geospatial and spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12978–12988.

Jieun Kim, Yujin Jeong, and Sung-Bae Cho. 2026a. Visual-linguistic abductive reasoning with llms for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2026, Rabat, Morocco, March 24-29, 2026*, Findings of ACL, pages 6529–6544. Association for Computational Linguistics.

Yoonji Kim, Jieun Kim, Yujin Jeong, and Sung-Bae Cho. 2026b. Diagnosing spatial consistency across perspectives and viewpoints in large vision-language models. In *The 64th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Roberta L. Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research (TMLR)*.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. 2025b. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *Preprint*, arXiv:2505.21500.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. 2026. Spatial-ladder: Progressive training for spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2026. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wufei Ma and 1 others. 2025. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Edvard I. Moser, Emilio Kropff, and May-Britt Moser. 2008. Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31:69–89.

John O'Keefe and Lynn Nadel. 1978. *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press, Oxford, UK.

Zhenyu Pan and Han Liu. 2026. Metaspatial: Reinforcing 3d spatial reasoning in VLMs for the metaverse. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Alexander W Siegel and Sheldon H White. 1975. The development of spatial representations of large-scale environments. In *Advances in child development and behavior*, volume 10, pages 9–55. Elsevier.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. 2025. Mind the gap: Benchmarking spatial reasoning in vision-language models. In *Greeks in AI Symposium 2025*.

Xingjian Tao, Yiwei Wang, Yujun Cai, Yifan Song, and Jing Tang. 2026. Viewfusion: Structured spatial thinking chains for multi-view reasoning. *arXiv preprint arXiv:2603.06024*.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, and 1 others. 2025. RoboBrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peiyao Wang and Haibin Ling. 2025. Svqa-r1: Reinforcing spatial reasoning in MLLMs via view-consistent reward optimization. *arXiv preprint arXiv:2506.01371*.

R. F. Wang and E. S. Spelke. 2002. Human spatial representation: Insights from animals. *Trends in Cognitive Sciences*, 6(9):376–382.

Ranxiao Frances Wang and Elizabeth S. Spelke. 2000. Updating egocentric representations in human navigation. *Cognition*, 77(3):215–250.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

Qiyao Xue, Weichen Liu, Shiqi Wang, Haoming Wang, Yuyang Wu, and Wei Gao. 2025. Reasoning path and latent state analysis for multi-view visual spatial

reasoning: A cognitive science perspective. *arXiv preprint arXiv:2512.02340*.

Jihan Yang, Shusheng Ang, Zizhe Li, Jinpeng Wang, Ruimao Gao, and Xiangyu Zhu. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. 2026. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations (ICLR)*.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. 2025. Spatial mental modeling from limited views. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, and 1 others. 2026. Mindcube: Spatial mental modeling from limited views. In *International Conference on Learning Representations (ICLR)*.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 2025. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv:2508.18265*.

A Experiments Settings

A.1 Training Settings

We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025), a 7B-parameter vision-language model, as the base model for all training stages. All stages employ LoRA (Hu et al., 2022) with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. Training is conducted on 4× NVIDIA A100 80GB GPUs with an effective batch size of 16 (per-device batch 1, gradient accumulation 4).

Stage 1 (Egocentric Perception SFT). We finetune on 4,240 egocentric spatial QA samples for 3 epochs with learning rate 2×10^{-4}. Training completes in approximately 1 hour.

Stage 2 (Egocentric-to-Allocentric Transformation SFT). Starting from the Stage-1 checkpoint, we train on 788 perspective-transformation samples for 3 epochs with learning rate 2×10^{-4}. Training completes in approximately 7 hours, reflecting the increased computational cost of 4-image multi-view inputs and longer chain-of-thought outputs.

Stage 3 (Spatial Consistency GRPO). Starting from the Stage-2 checkpoint, we apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024) on 5,516 consistency-reward samples for 3 epochs. We use learning rate 5×10^{-6}, clipping $\epsilon = 0.2$, KL coefficient $\beta = 0.04$, rollout temperature 0.9, $G = 4$ generations per prompt. Training completes in approximately 21 hours.

A.2 Training Data

Data Construction and Statistics. To construct transformation training data, we build on SCOPE (Kim et al., 2026b), one of the few publicly available multi-view resources that provides the scene-level spatial metadata and camera pose information needed for egocentric-to-allocentric transformation learning. Using this metadata, we automatically generate reasoning traces without manual annotation. For each scene, we sample multi-view observations from different camera headings. Using the scene notes, we parse the subject, target, and spatial relation, and deterministically compute the target's image-region position and the subject's heading direction for each view based on the scene geometry. As a result, egocentric observations vary across camera viewpoints, whereas the allocentric relation remains invariant because all views correspond to the same underlying scene. We then instantiate these reasoning traces in natural language following the four-component structure defined above, yielding view-wise traces followed by a final allocentric answer.

Table 9 summarizes the training data used in each stage.

Table 9: **Statistics of training data** used across the three training stages.

Stage	# Samples	# Views
SFT-P	4,240	1
SFT-T	788	4
GRPO-C	5,516	2/4/6/8

A.3 Clarification of Reported Improvement Metrics

We clarify that the +3.69%p (multi-view) and +4.72%p (general spatial) gains stated in the abstract and introduction are averaged improvements over the strongest prior SoTA per benchmark (MindCube: +4.78%p vs. Spatial Reasoner; MMSI: +1.70%p vs. VILASR-7B; VisualSpatial: +4.60%p vs. Spatial-SSRL; general spatial: vs. Spatial-Thinker), whereas Δ in tables denotes improvement over the Qwen2.5-VL-7B backbone.

Licenses, Terms of Use, and Intended Use. We use all public datasets, benchmarks, pretrained models, and software artifacts in accordance with their original licenses and terms of use throughout data construction, model training, and evaluation. SCOPE (Kim et al., 2026b) and other public benchmarks are used only for their intended research purposes, including data construction, model training, and evaluation. Our generated training traces are derived from available scene metadata and do not introduce personally identifiable information or human-subject data. Any newly created or derivative artifacts, including the trained model checkpoints, training data, and code, will be released for research use under appropriate licenses, together with documentation of their intended use and conditions compatible with the original resources.

Artifact Documentation. The artifacts in this work are designed for visual spatial reasoning, with a focus on object-centric spatial relations under viewpoint variation. All generated questions, reasoning traces, and answers are in English. The data covers egocentric observations, allocentric spatial relations, and egocentric-to-allocentric transformations across multi-view settings. The

artifacts do not target demographic attributes or human-centered linguistic phenomena, and no demographic group annotations are used.

Benchmark Statistics. Table 10 summarizes the evaluation benchmarks used in our experiments. The benchmark suite covers a diverse range of spatial relations beyond basic directional relations such as left/right/front/behind. MindCube includes directional relations, diagonal movement, rotation, distance, and facing relations, while MMSI covers directional, compass-based, height, and distance relations. ViewSpatial-Bench evaluates 18 directional relation tokens in 3D space. What's Up and SPAR-Bench focus on object-level spatial reasoning involving relative positions and spatial configurations, and BLINK further covers containment, relative position, distance, facing, and other spatial interactions. Overall, the benchmarks span both directional and non-directional spatial reasoning patterns.

A.4 Training Prompt Templates

We construct three training datasets across the pipeline. Below we detail the input/output formats and prompt templates used in each stage.

Stage 1: Perception SFT. The perception dataset consists of 4,240 single-image, camera-centric spatial reasoning samples generated from the SCOPE perception set. After excluding eight scenes overlapping with MindCube (one from DL3DV-10K and seven from WildRGB-D), we construct QA pairs from 265 scenes using eight viewpoints and two QA pairs per viewpoint. Each sample is automatically generated from scene metadata, including object locations and camera poses, to derive spatial relations. A deduplication analysis further found no overlap with MMSI-Bench or ViewSpatial-Bench. Each sample follows the template:

Egocentric Perspective Input Template

```
[IMAGE]
From the camera's perspective, where is {object_subject} relative to {object_reference}?
```

The ground-truth answer is a single directional label from {*front*, *behind*, *left*, *right*, *front-left*, *front-right*, *behind-left*, *behind-right*}.

Stage 2: Transformation. Stage 2 consists of two phases sharing the same task formulation: supervised fine-tuning (SFT) on structured reasoning traces, followed by GRPO-based consistency training on variable-view inputs.

Stage 2-1: Transformation SFT. The transformation SFT dataset consists of 788 samples generated from SCOPE (Blender-rendered scenes) (Kim et al., 2026b). Each scene is rendered from 8 viewpoints, producing 2 samples of 4-view inputs per scene. The input template is:

Transformation Input Template

```
View 1: [IMAGE]
View 2: [IMAGE]
View 3: [IMAGE]
View 4: [IMAGE]
The images show the same scene from 4 different camera viewpoints ({angles}).
Determine the spatial relation of the {target} with respect to the {subject}'s facing direction.
For each view, describe (1) where the {target} appears in the image, (2) the facing direction of the {subject} in the image, and (3) the spatial relation of the {target} from the {subject}'s perspective. Then conclude with the consistent allocentric relation across all views.
```

The ground-truth response follows a structured chain-of-thought format:

Transformation Output Template

```
View 1:
The {target} appears in {region}.
The {subject} is facing {heading}.
Therefore, the {target} is {relation} the {subject}.

View 2: ···

Across all views, the spatial relation remains consistent.

Final Answer: {relation}
```

Stage 2-2: GRPO Consistency Training. The GRPO dataset consists of 5,516 samples drawn from the same scenes as Stage 2-1, using 14 distinct view-count subsets with $N \in \{2, 4, 6, 8\}$ views per sample. The input template generalizes Stage 2-1 to variable view counts:

Table 10: **Benchmark Datasets Used for Evaluation.**

Benchmark	Reference	Data Size	Eval Size	Task Focus
In-domain Benchmarks				
ViewSpatial-Bench	Li et al. (2025b)	5,711 questions, 5,250 images	1,000 questions(seed 55)	Multi-perspective spatial localization
MMSI-Bench	Yang et al. (2026)	1,000 questions, 120,000+ images	1,000 questions	Multi-image spatial intelligence
MindCube	Yin et al. (2026)	21,154 questions, 3,268 images	1,200 questions	Spatial mental modeling
Out-of-domain Benchmarks				
What'sUp	Kamath et al. (2023)	820 images, 820 questions	820 questions	2D spatial relations
BLINK	Fu et al. (2024)	3,807 questions, 7,358 images	400 questions(Spatial Only)	Visual perception
SPAR-Bench	Zhang et al. (2025)	7,210 questions	7,210 questions	Multi-task spatial reasoning

GRPO Input Template

```
View 1: [IMAGE]
⋮
View N: [IMAGE]
The images show the same scene from N
different camera viewpoints ({angles}).
Determine the spatial relation of the
{target} with respect to the {subject}'s
facing direction.
For each view, describe (1) where the
object appears, (2) the {subject}'s facing
direction, and (3) the transformed relation.
End with 'Final Answer: [relation]'.
```

Unlike Stage 2-1, no ground-truth answer is provided; the model is trained purely via GRPO reward signals based on cross-view consistency.

B Additional Analysis

B.1 Generalization to Different Model Scales

Our main experiments are conducted with Qwen2.5-VL-7B as the backbone. To examine whether the performance gains stem from the training methodology itself rather than model capacity, we apply our pipeline to a smaller backbone, Qwen2.5-VL-3B, without any modification to the training procedure.

As shown in Table 11, SCoRE (3B) consistently outperforms both the base model and the Ego+Alloc SFT baseline across allocentric subtasks, achieving the highest Alloc avg. and Overall scores at the 3B scale. This confirms that our two-stage training strategy, egocentric perception followed by allocentric transformation, yields consistent improvements regardless of model scale, validating the generality of our approach.

B.2 Prompting Analysis for Spatial Reasoning

We conduct a pilot study to examine whether existing LVLMs can perform egocentric-to-allocentric spatial transformation using prompting alone. Following the allocentric-integration task setting and prompt structure of SCOPE (Kim et al., 2026b), we evaluate five LVLMs: RoboBrain-7B (Team et al., 2025), Gemma3-4B (Gemma Team et al., 2025), LLaVA-OneVision-8B (Li et al., 2025a), InternVL2.5-8B (Zhu et al., 2025), and Llama-3.2-11B-Vision (Grattafiori et al., 2024).

In the direct prompting setting, the model is asked to infer the final allocentric relation from multi-view egocentric observations without explicit intermediate transformation steps. In the CoT prompting setting, the prompt encourages the model to first reason about each view-specific egocentric observation and then integrate them into a final allocentric relation. The ground-truth answer is a single directional label from {*front*, *behind*, *left*, *right*, *front-left*, *front-right*, *behind-left*, *behind-right*}.

As shown in Figure 2, egocentric-to-allocentric CoT prompting improves performance for several LVLMs, suggesting that explicit transformation cues can help models reason beyond viewpoint-specific observations. However, the effect is not uniformly positive: some models show performance drops under prompting, such as InternVL2.5-38B (62.77% $\rightarrow$ 58.80%) and Qwen2.5-VL-3B (56.2% $\rightarrow$ 54.90%). These mixed results indicate that prompting alone does not reliably induce stable egocentric-to-allocentric transformation. This motivates our training-based formulation, which explicitly supervises the transformation process and further refines spatial consistency through GRPO.

B.3 Relational Consistency Detail

B.3.1 Dataset Construction Pipeline

The dataset is automatically generated from MindCube-tiny through a multi-stage pipeline for constructing grounded spatial reasoning tasks. Although the original objective is to generate cognitive maps in JSON format, we reformulate this process as a question-answering task. Rather than directly predicting the full cognitive map, each

Table 11: **Generalization to Different Model Scales** on ViewSpatial-Bench. Best per scale is **bolded**.

Model	Egocentric		Allocentric			Ego avg.	Alloc avg.
	Rel.Dir.	Obj.Ori.	Obj.Ori.	Rel.Dir.	Sce.Sim.		
3B Scale							
Qwen2.5-VL-3B	36.04	19.68	34.34	30.76	18.91	27.86	28.00
Ego+Alloc SFT	29.67	16.58	36.31	25.00	25.00	23.12	28.77
SCoRE (Ours)	**32.00**	**21.24**	**42.26**	**26.97**	**22.73**	**26.62**	**30.65**
7B Scale							
Qwen2.5-VL-7B	40.67	27.98	38.69	42.76	24.06	34.32	35.17
Ego+Alloc SFT	42.33	23.83	30.95	38.16	28.34	33.08	32.48
SCoRE (Ours)	**41.67**	**32.64**	**57.74**	**48.03**	**33.16**	**37.15**	**46.31**

scene is first converted into a structured cognitive map representation, from which we derive multiple types of question-answer pairs that target distinct spatial reasoning abilities. These QA pairs serve as intermediate supervision, and the resulting model responses are later aggregated and transformed into the final JSON-formatted cognitive map.

The generation process begins by iterating over the raw samples and selecting scenes that contain the spatial relation categories `among`, `around`, and `rotation`. These relation types contain rich geometric and viewpoint-aware spatial information, making them suitable for constructing grounded spatial reasoning tasks. Each filtered scene is converted into the grounded cognitive map representation used in MindCube evaluation. We use this representation not as a direct prediction target, but as a source of reliable ground-truth spatial information. The cognitive map represents each scene on a discrete 10×10 grid, explicitly encoding object identities, object coordinates, viewer positions, and viewer orientations. Based on this structured representation, we derive ground-truth answers for both allocentric object-to-object reasoning and egocentric viewpoint reasoning. After constructing the cognitive maps, the pipeline generates two categories of spatial reasoning tasks. The first category targets object-to-object relational reasoning. These questions evaluate whether the model can infer directional relations, compare relative distances, and determine spatial ordering between objects with respect to a given reference view. Directional relation questions are constructed by computing the relative position of one object with respect to another in the local coordinate frame of the specified view. The following template is used for directional relation reasoning:

Directional Relation Template

```
[IMAGE]
Referring to {view_name}, where is {object_subject} relative to {object_reference}?
```

The corresponding answers are directional labels such as `left`, `right`, `front`, and `behind`.

Distance comparison questions are generated by selecting the two nearest objects relative to an anchor object and asking the model to identify which object is closer.

Distance Comparison Template

```
[IMAGE]
Referring to {view_name}, which is closer to {anchor_object}: {object_a} or {object_b}?
```

Spatial ordering tasks are generated by projecting object coordinates into the local coordinate frame of the viewer and sorting objects along the horizontal axis.

Spatial Ordering Template

```
[IMAGE]
Referring to {view_name}, order these objects from your left to right: {object_list}
```

The second category focuses on egocentric viewpoint reasoning. These questions require the model to infer which object appears in a specific direction relative to the viewer's current orientation.

For each viewpoint, the pipeline computes whether objects are located in front of, behind, left of, or right of the observer under the viewer-centered coordinate system. The nearest valid

object in the queried direction is selected as the ground-truth answer.

The primary egocentric reasoning template is shown below:

Egocentric Perspective Input Template

```
[IMAGE]
From the perspective of {view_name}, what
object is to your {direction}?
```

The candidate directions include `front`, `behind`, `left`, and `right`. If no object exists in the queried direction, the answer is assigned as `none`.

During evaluation, the model's QA predictions are converted back into a cognitive map. Each predicted answer is treated as a spatial constraint: object-to-object relation predictions provide pairwise constraints between objects, while viewpoint predictions provide constraints between objects and the corresponding view positions. These constraints are aggregated within each source scene to estimate object coordinates on the same 10×10 grid. The reconstructed cognitive map is then compared with the ground-truth cognitive map using the MindCube cognitive map similarity metric, allowing us to evaluate whether the QA predictions form a spatially consistent JSON-formatted cognitive map.

B.3.2 Relation Consistency Evaluation

We evaluate whether answers to multiple relation direction questions are mutually consistent. Each predicted direction $\hat{d}$ is first converted from the reference view coordinate frame into a global offset:

$$(dx, dy) = \mathrm{LocalToGlobal}(f_v, \hat{d}), \tag{14}$$

where f_v denotes the orientation of the reference view. The resulting (dx, dy) indicates whether the prediction supports a horizontal or vertical relation in the global coordinate system.

For each object pair (A, B), we aggregate all predicted relations as votes along the x- and y-axes:

$$xv_{AB} = \sum_i \Delta x_{AB}^{(i)}, \tag{15}$$

$$yv_{AB} = \sum_i \Delta y_{AB}^{(i)}, \tag{16}$$

where $\Delta x_{AB}^{(i)}$ and $\Delta y_{AB}^{(i)}$ are the global directional votes contributed by the i-th QA prediction for the pair (A, B).

The ground-truth relative displacement of the same pair is computed from the cognitive map coordinates:

$$gx_{AB} = x_A - x_B, \tag{17}$$

$$gy_{AB} = y_A - y_B. \tag{18}$$

Here, gx_{AB} and gy_{AB} indicate the true horizontal and vertical offsets between objects A and B.

We then select the axis with stronger accumulated evidence:

$$a_{AB} = \begin{cases} x, & |xv_{AB}| \geq |yv_{AB}|, \\ y, & |yv_{AB}| > |xv_{AB}|. \end{cases} \tag{19}$$

Based on the selected axis, the predicted sign is defined as:

$$\hat{s}_{AB} = \begin{cases} \mathrm{sign}(xv_{AB}), & a_{AB} = x, \\ \mathrm{sign}(yv_{AB}), & a_{AB} = y. \end{cases} \tag{20}$$

The corresponding ground-truth sign is:

$$g_{AB} = \begin{cases} \mathrm{sign}(gx_{AB}), & a_{AB} = x, \\ \mathrm{sign}(gy_{AB}), & a_{AB} = y. \end{cases} \tag{21}$$

A pair is counted as correct if the predicted sign matches the ground-truth sign:

$$m_{AB} = \mathbb{1}[\hat{s}_{AB} = g_{AB}]. \tag{22}$$

Pairs are excluded when no valid vote is available or when the ground-truth displacement on the selected axis is zero.

Relation Accuracy. Relation Accuracy measures the fraction of evaluated object pairs whose aggregated relation prediction agrees with the ground-truth cognitive map:

$$\mathrm{RA} = \frac{\sum_{(A,B)\in\mathcal{P}_{\mathrm{eval}}} m_{AB}}{|\mathcal{P}_{\mathrm{eval}}|}, \tag{23}$$

where $\mathcal{P}_{\mathrm{eval}}$ is the set of valid object pairs after excluding pairs with no vote or zero ground-truth displacement.

Pair Coverage. Pair Coverage measures the proportion of all possible object pairs that are included in the evaluation:

$$\mathrm{PC} = \frac{|\mathcal{P}_{\mathrm{eval}}|}{\binom{|\mathcal{O}|}{2}}, \tag{24}$$

where $\mathcal{O}$ denotes the object set in the scene.

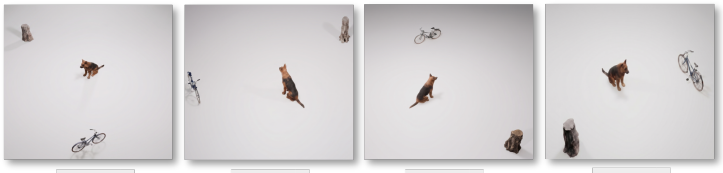
Global Consistency. Global Consistency measures whether the entire scene is spatially consistent. A scene is considered globally consistent only if all evaluated object pairs are correct:

$$\mathrm{GC} = \mathbb{1}[\mathrm{RA} = 1]. \tag{25}$$

Thus, Relation Accuracy evaluates pair-level correctness, Pair Coverage measures the evaluated relation density, and Global Consistency captures scene-level spatial coherence.

B.4 Training Data Provenance

We clarify the source and construction procedure of the training data used in each stage of SCoRE. The Stage-2 SFT and GRPO training data are constructed from Blender-rendered scenes provided by SCOPE. The Stage-1 SFT-P data, consisting of 4,240 QA samples, are automatically generated from the SCOPE perception set rather than sourced from any external QA dataset: after removing 8 scenes that overlap with MindCube (1 from DL3DV-10K and 7 from WildRGB-D), we construct QA pairs from the remaining 265 scenes (8 viewpoints $\times$ 2 QA pairs per viewpoint). Each QA sample is generated automatically from scene metadata, including object locations and camera poses, to derive the corresponding spatial relations, without manual annotation.



Egocentric-to-Allocentric Transformation

User Instruction:
The images show the same scene from 4 different camera viewpoints.
Determine the spatial relation of the bicycle with respect to the dog's facing direction.
For each view, describe (1) where the bicycle appears in the image, (2) the facing direction of the dog in the image, and (3) the spatial relation of the bicycle from the dog's perspective.

Target Response:

View 1:
The bicycle appears in the lower-center region of the image.
The dog appears near the center of the image and is facing diagonally toward-left.
From the dog's perspective, its front direction points diagonally toward the lower-left side of the image, so the bicycle in the lower-center region lies toward the dog's front-left region.
Therefore, the bicycle is to the front-left of the dog.

View 2:
The bicycle appears on the left side of the image.
The dog is facing diagonally away-left relative to the camera.
From the dog's perspective, its front direction points diagonally away-left, so the bicycle on the image-left side lies toward the dog's front-left region.
Therefore, the bicycle is to the front-left of the dog.

View 3:
The bicycle appears in the upper-left region of the image.
The dog appears on the right side of the image and is facing away from the camera.
From the dog's perspective, its front direction points away from the camera, so the bicycle in the upper-left region lies toward the dog's front-left region.
Therefore, the bicycle is to the front-left of the dog.

View 4:
The bicycle appears on the right side of the image.
The dog is facing diagonally toward-right relative to the camera.
From the dog's perspective, its front direction points diagonally toward the right side of the image, so the bicycle on the image-right side lies toward the dog's front-left region.
Therefore, the bicycle is to the front-left of the dog.

Across all views, the spatial relation remains consistent.

Final Answer: front-left

Figure 8: **Representative SFT training example.** The example shows the user instruction and target response used to supervise egocentric-to-allocentric transformation across four camera viewpoints.