Depth as Verifiable Geometric Feedback: Enhancing Spatial Reasoning in Multimodal Large Language Models

Yepeng Zhang* **Ruohang Xu*** **Zihao Yue** **Boshen Xu** **Qin Jin†**
AIM3 Lab, School of Information, Renmin University of China
{yepzhang,johan,qjin}@ruc.edu.cn

Abstract

Multimodal large language models (MLLMs) still struggle with spatial reasoning, particularly on tasks requiring consistent geometric understanding across multiple image regions. While recent works incorporate depth supervision into MLLMs, it remains unclear what forms of geometric learning are most beneficial for spatial reasoning transfer. In this work, we propose **Multi-point Geometric Consistency Alignment** (MGCA), a geometry-guided post-training framework that uses depth annotations as *verifiable geometric feedback* without requiring depth input at inference time. MGCA samples multiple image regions and optimizes the model through a unified geometric reward that jointly captures *metric grounding* and *relative geometric consistency*. Specifically, the reward combines a metric depth objective with a margin-aware ranking objective and is optimized via Group Relative Policy Optimization (GRPO). Experiments on multiple spatial reasoning benchmarks show that MGCA consistently improves spatial understanding and outperforms comparable open-weight MLLMs of similar scale. More importantly, our analysis reveals that optimizing relative ordering alone can achieve high ranking accuracy while failing to learn transferable geometric representations due to the lack of metric grounding. In contrast, jointly optimizing metric depth and relational consistency leads to substantially stronger spatial reasoning performance. These findings suggest that effective spatial alignment in MLLMs requires metrically grounded relational learning, rather than relation-only supervision.

1 Introduction

Multimodal large language models (MLLMs) have recently achieved remarkable progress across a wide range of vision-language tasks, including image captioning, visual question answering, and multimodal dialogue (Radford et al., 2021; Li et al., 2023; Liu et al., 2023; Alayrac et al., 2022; Bai et al., 2025). However, spatial reasoning remains a major weakness (Liu et al., 2025; Wang et al., 2024; Ma et al., 2025; Wang et al., 2025b). While current MLLMs can often recognize objects and describe scenes fluently, they frequently fail on questions requiring consistent geometric understanding, such as comparing distances, reasoning about depth and occlusion, or inferring spatial relations among multiple regions in an image (Wang et al., 2024; Ma et al., 2025; Wang et al., 2025b). In such cases, models may produce answers that are linguistically plausible yet geometrically inconsistent, suggesting that current visual-language alignment remains weakly grounded in the underlying 3D structure of the physical world.

Depth provides a natural source of geometric supervision for addressing this limitation. As an intermediate representation between 2D observations and 3D scene structure, depth encodes essential spatial cues including distance, scale, perspective, and occlusion ordering (Saxena et al., 2008; Eigen et al., 2014; Hoiem et al., 2011). Recent studies have explored incorporating geometry into vision-language systems through depth-map inputs, spatial instruction tuning, 3D region annotations, or language-based depth prediction objectives (Chen et al., 2024; Cheng et al., 2024; Chen et al., 2026; Cai et al., 2025; Yu et al., 2026). These approaches suggest that geometry-aware supervision can improve spatial perception beyond standard image-text alignment.

Nevertheless, an important question remains insufficiently understood: *what aspects of depth supervision actually improve spatial reasoning in MLLMs?* A straightforward solution is to optimize models for metric depth prediction (Cai et al., 2025; Yu et al., 2026). However, accurate depth estimation alone does not necessarily translate into strong spatial reasoning ability. Many spatial reasoning

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 47258–47274
October 24-29, 2026 ©2026 Association for Computational Linguistics

tasks are inherently relational: they ask which object is closer, whether one region is behind another, or whether a spatial configuration is geometrically consistent (Wang et al., 2024; Ma et al., 2025; Wang et al., 2025b; Cheng et al., 2024). Conversely, optimizing only relative depth ordering is also insufficient. Pure ranking supervision lacks metric grounding and may encourage scale-free relational shortcuts that achieve high ranking accuracy without learning physically meaningful geometry (Zeng et al., 2024; Bhat et al., 2023; Zoran et al., 2015). As a result, models trained only on relative ordering may perform well on proxy ranking objectives while transferring poorly to downstream spatial reasoning tasks.

These observations motivate a different perspective: rather than treating depth as an auxiliary modality or a standalone prediction target (Cheng et al., 2024; Chen et al., 2026; Cai et al., 2025; Chen et al., 2024), we view depth as a source of *verifiable geometric feedback* for reasoning-oriented alignment. From this perspective, the key objective is not to make MLLMs become dense depth estimators, but to align their internal representations with geometrically consistent spatial reasoning behavior.

To this end, we propose **Multi-point Geometric Consistency Alignment** (MGCA), a geometry-guided post-training framework for improving spatial reasoning in MLLMs. Given a depth-annotated image, MGCA samples multiple local regions with reliable and sufficiently separated depths, and prompts the model to predict their depths in a structured textual format. The generated predictions are then evaluated using a unified geometric consistency reward. The reward contains two complementary components: (1) a *metric depth reward* that encourages metrically plausible geometric understanding, and (2) a *margin-aware ranking reward* that enforces consistent relative depth relations across regions. To stabilize optimization across samples with varying difficulty and depth scales, we further incorporate online reward normalization during reinforcement learning (Van Hasselt et al., 2016). The entire framework is optimized using Group Relative Policy Optimization (GRPO), enabling geometry-guided post-training through fully verifiable supervision (Shao et al., 2024).

We evaluate MGCA on multiple spatial reasoning benchmarks, including Spatial457 (Wang et al., 2025b), SpatialEval (Wang et al., 2024), and 3DSR-Bench (Ma et al., 2025). Experimental results show that MGCA consistently improves spatial reasoning performance over baselines and MGCA variants using only metric or only ranking supervision. More importantly, our controlled analyses reveal a key insight: metric grounding and relative geometric reasoning play fundamentally complementary roles. Relative ranking alone can produce strong proxy-level ordering accuracy while failing to learn transferable geometric representations, whereas metric grounding provides physically meaningful spatial calibration that substantially improves downstream transfer. Better spatial reasoning performance emerges only when both signals are jointly optimized.

In summary, our main contributions in this work are as follows:

- We propose MGCA, a geometry-guided post-training framework that uses depth annotations as verifiable geometric feedback for improving spatial reasoning in MLLMs, without requiring depth input at inference time.
- We design a unified geometric consistency reward that jointly combines metric depth grounding, margin-aware relational reasoning, and online reward normalization for stable alignment.
- We provide systematic empirical and analytical evidence that effective spatial reasoning requires the joint optimization of metric geometry and relational consistency, rather than either signal alone.

2 Related Work

2.1 Benchmarks for Spatial Understanding

Recent benchmarks have revealed persistent limitations in the spatial reasoning abilities of MLLMs (Liu et al., 2025; Xu et al., 2024; Zheng et al., 2025; Ye et al., 2025; Zheng et al., 2026). Spatial457 (Wang et al., 2025b) evaluates structured spatial reasoning in synthetic scenes, including localization, orientation, occlusion, and collision-related judgments. SpatialEval (Wang et al., 2024) focuses on spatial relations and layout-level reasoning across maps, grids, mazes, and real images. 3DSRBench (Ma et al., 2025) targets 3D spatial reasoning in natural images through manually annotated questions involving relative positions, distances, and viewpoints. Beyond static-image evaluation, recent benchmarks such as VSI-Bench (Yang et al., 2025a), VSI-SUPER(Yang et al., 2025b) and OmniSpatial (Jia et al., 2025) further extend spatial reasoning assessment to video-

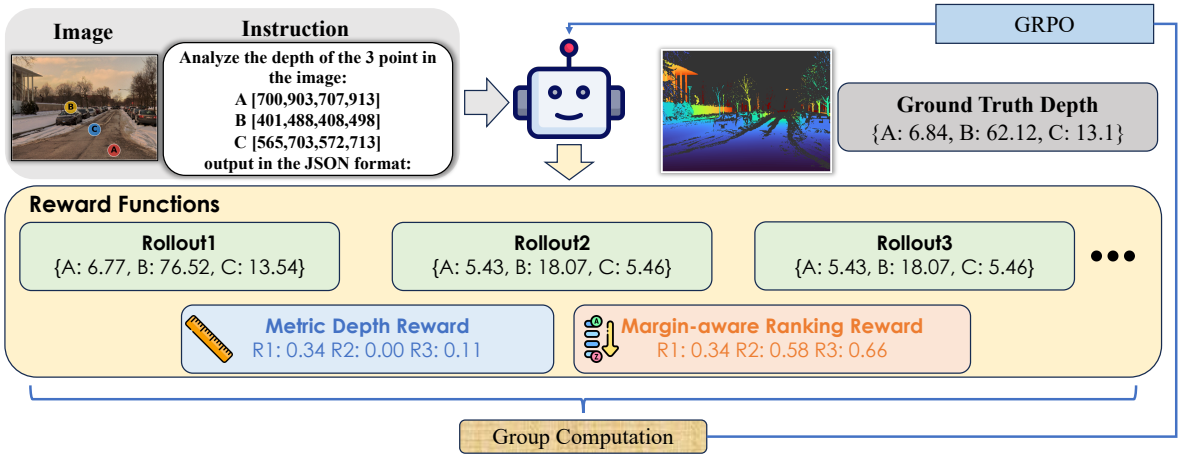


Figure 1: Overview of MGCA. Given a depth-annotated image, MGCA constructs a structured three-region depth-prediction task and evaluates generated responses with metric and relative geometric rewards for GRPO-based post-training. Depth annotations are not required at inference time.

based memory and broader spatial interaction settings.

We evaluate MGCA on Spatial457, SpatialEval, and 3DSRBench, which provide complementary coverage of structured geometric reasoning, global spatial organization, and 3D relations in natural imagery.

2.2 Geometric Supervision for Spatial Reasoning

Recent studies improve spatial reasoning in MLLMs by incorporating explicit geometric supervision. SpatialVLM (Chen et al., 2024) constructs large-scale quantitative spatial question–answer data from automatically generated geometric annotations, while SD-VLM (Chen et al., 2026) injects depth information into visual representations through depth positional encoding. DepthVLM (Yu et al., 2026) jointly supports language generation and dense metric depth prediction, and SpatialThinker (Batra et al., 2025) uses structured scene-graph grounding with reinforcement learning for multi-step 3D reasoning.

MGCA takes a different approach: it does not require additional depth input at inference time or train the model to output dense depth maps. Instead, it converts the metric depths and relative geometry of sampled local regions into verifiable rewards for post-training, using depth as an alignment signal while preserving the original image–text inference interface.

3 Method

MGCA is a geometry-guided post-training method for improving spatial reasoning in MLLMs. It uses depth annotations only during post-training and requires no additional depth or geometric input at inference time. Rather than treating depth prediction as the final objective, MGCA converts local depth annotations into verifiable geometric feedback for reasoning-oriented alignment.

Figure 1 illustrates the overall pipeline. Given a depth-annotated image, MGCA samples three reliable local regions with sufficiently separated depths and constructs a structured depth-prediction instruction. For each image-instruction pair, the model generates multiple candidate responses. Each response is scored by a geometric consistency reward that evaluates both metric depth plausibility and relative geometric consistency across the sampled regions. The resulting rewards are used for GRPO-based post-training (Shao et al., 2024).

3.1 Multi-point Geometric Supervision

MGCA constructs a spatially structured supervision instance from each depth-annotated image. Rather than asking the model to predict the depth of an isolated region, we sample three local regions whose depths are individually reliable and mutually distinguishable. This construction provides both metric targets for individual regions and relative geometric constraints within the same scene.

Reliable triplet construction. Given an RGB image I and its associated depth map D, we sample local candidate regions from valid portions of

the image. For a region p_i, its reference depth is computed as the mean depth over valid pixels:

$$g_i = \frac{1}{|\Omega_i|} \sum_{u \in \Omega_i} D(u), \tag{1}$$

where Ω_i denotes the set of valid depth pixels within p_i. We discard regions with insufficient valid-depth coverage or excessive internal depth variation, since their region-level depths are unreliable near missing areas or depth discontinuities. We then select three non-overlapping regions from the remaining candidates. Let their sorted reference depths be $g_{(1)} < g_{(2)} < g_{(3)}$. To avoid ambiguous relative relations, we retain a triplet only when

$$\frac{g_{(2)}}{g_{(1)}} \geq \tau_{\text{sep}}, \qquad \frac{g_{(3)}}{g_{(2)}} \geq \tau_{\text{sep}}. \tag{2}$$

The resulting triplet contains three stable local depth targets together with clear near-far relations.

Region-specified depth prediction. The three sampled regions are represented in the instruction by their normalized bounding-box coordinates and associated with identifiers A, B, and C. Given the original RGB image and the region-specified instruction x, the model is asked to predict one metric depth value for each region in a fixed textual format, such as `A: <depth>, B: <depth>, C: <depth>`. The generated response y is parsed into

$$\hat{\mathbf{d}} = \text{Parse}(y) = (\hat{d}_1, \hat{d}_2, \hat{d}_3), \qquad \mathbf{g} = (g_1, g_2, g_3). \tag{3}$$

A response is valid only if all three predicted depths can be successfully parsed from the required output format. The colored region overlays in Figure 1 are provided only for visualization and are not part of the model input.

The purpose of using three regions is not simply to provide more depth targets. A triplet also defines pairwise depth ordering and relative depth-gap structure, allowing the model response to be evaluated for cross-region geometric consistency. MGCA therefore turns depth annotations into supervision for relational spatial reasoning without requiring manually annotated spatial questions.

3.2 Geometric Consistency Reward

The structured prediction task allows each response to be verified directly against the reference depths of the sampled regions. MGCA evaluates a response from two complementary aspects: whether the predicted values are metrically plausible, and whether they preserve the relative geometry among regions. The final reward is defined as

$$R(I, x, y) = r_{\text{fmt}}(y) \left[w_d \widetilde{R}_{\text{depth}} + w_r \widetilde{R}_{\text{rank}} \right], \tag{4}$$

where $r_{\text{fmt}}(y) \in \{0, 1\}$ indicates whether all three predicted depths can be parsed from the required output format. R_{depth} measures metric depth accuracy, while R_{rank} measures relative geometric consistency among the sampled regions.

Metric depth reward. For each sampled region, we measure the discrepancy between the predicted depth $\hat{d}_i$ and the reference depth g_i using a symmetric relative error and a clipped absolute hinge error:

$$e_i^{\text{rel}} = \frac{2|\hat{d}_i - g_i|}{|\hat{d}_i| + |g_i|}, \tag{5}$$

$$e_i^{\text{abs}} = \min\Big(\max\Big(|\hat{d}_i - g_i| - \tau_{\text{abs}},\, 0 \Big),\, c_{\text{abs}} \Big). \tag{6}$$

The metric depth reward is then averaged over the three regions:

$$R_{\text{depth}} = \frac{1}{3} \sum_{i=1}^{3} \exp\Big(-\alpha_{\text{rel}} e_i^{\text{rel}} - \beta_{\text{abs}} e_i^{\text{abs}} \Big). \tag{7}$$

The relative term accommodates variation in scene depth ranges, while the absolute term penalizes predictions with large metric deviations.

Margin-aware ranking reward. The metric reward evaluates individual depth predictions, but does not explicitly constrain the relative geometry among the sampled regions. For the three region pairs in $\mathcal{E} = \{(1,2), (1,3), (2,3)\}$, we compute an ordering score s_{ij}^{ord} and a gap-consistency score s_{ij}^{gap}. The former rewards the correct near–far relation, while the latter measures whether the predicted depth difference agrees with the reference margin. Pairs with nearly identical reference depths are assigned smaller weights, and excessively large margins are clipped to avoid dominating the reward.

We aggregate the pairwise scores into an ordering reward R_{ord} and a gap-consistency reward R_{gap}, and define

$$R_{\text{rank}} = \lambda R_{\text{ord}} + (1 - \lambda) R_{\text{gap}}. \tag{8}$$

This reward encourages relative geometric consistency by accounting for both depth ordering and

Model	# Params	3DSRBench	SpatialEval	Spatial457
Open-weight MLLMs				
Qwen3-VL-Instruct (Bai et al., 2025)	4B	56.25	60.65	37.91
InternVL3.5 (Wang et al., 2025a)	4B	52.62	48.89	34.30
LLaVA-OneVision-1.5 (An et al., 2025)	4B	53.11	56.91	36.86
Proprietary MLLMs				
Gemini 3.1 Flash-Lite (Team et al., 2023)	N/A	60.63	80.09	38.86
GPT-5 mini (Welsby and Cheung, 2023)	N/A	**63.91**	**90.66**	34.14
Geometry-guided Post-training				
Qwen3-VL-Instruct + MGCA	4B	57.08	63.15	**42.01**

Table 1: Comparison with representative MLLMs on spatial understanding benchmarks. "N/A" denotes undisclosed parameter counts. MGCA is applied to the Qwen3-VL-Instruct backbone and requires no additional depth input at inference time.

pairwise margin structure among the sampled regions. The detailed scoring and aggregation functions are provided in Appendix B.

Component-wise online normalization. The metric depth and ranking rewards may have different value distributions during training. To balance their contributions, we maintain online running statistics for each reward component and normalize them separately before aggregation:

$$z_c = \mathrm{clip}\left(\frac{R_c - \mu_c}{\sqrt{\sigma_c^2 + \varepsilon}}, -\gamma, \gamma\right), \qquad \widetilde{R}_c = \sigma(z_c), \qquad c \in \{\mathrm{depth}, \mathrm{rank}\}, \tag{9}$$

where μ_c and σ_c^2 are exponential-moving-average statistics computed from synchronized training batches. Normalization is activated after a short warm-up stage; before that, the raw reward components are used directly. This operation is applied before weighted reward aggregation and is separate from the group-relative normalization in GRPO.

3.3 GRPO-based Post-training

We optimize the model with GRPO (Shao et al., 2024). For each image–instruction pair, the policy samples a group of candidate responses. Each response is parsed into predicted depths and evaluated using the geometric consistency reward in Eq. 4. GRPO then favors responses that achieve higher geometric consistency relative to other candidates generated from the same input.

MGCA does not modify the GRPO objective. Instead, it provides multi-point supervision and a verifiable reward that jointly evaluates metric depth plausibility and relative geometric consistency. Depth annotations are used only to construct training instances and compute rewards during post-training; at inference time, the adapted MLLM follows the original image–text interface without additional depth input.

4 Experiments and Analysis

We evaluate MGCA on three spatial understanding benchmarks and analyze how different forms of depth-derived supervision transfer to downstream spatial reasoning. In addition to comparisons with representative MLLMs, we conduct controlled analyses on downstream capability groups and a depth-based proxy task.

4.1 Experimental Setup

Training data. We use DIODE (Vasiljevic et al., 2019) as the only source of depth supervision for post-training. Multi-point supervision instances are constructed from its training split by sampling three local regions that satisfy the validity and depth-separation constraints described in Section 3.1. The validation split is used only for proxy-task evaluation, and none of the downstream spatial understanding benchmarks is included in post-training.

Backbone and training. We use Qwen3-VL-4B-Instruct[1] (Bai et al., 2025) as the backbone model and perform parameter-efficient post-training with Low-Rank Adaptation (LoRA) (Hu et al., 2022). The model is optimized with GRPO using the proposed geometric consistency reward. Unless otherwise specified, MGCA uses $w_d = 0.3$ and $w_r = 0.7$. Implementation details, reward hyperparameters, and training configurations are provided in Appendix A.

Evaluation. We evaluate downstream spatial reasoning on Spatial457 (Wang et al., 2025b), SpatialEval (Wang et al., 2024), and 3DSRBench (Ma

[1] https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct

Model	Local Geo.	Relation	Global Org.
Base	42.89	44.75	60.65
Depth-only	43.97	46.01	59.78
Rank-only	41.22	43.94	54.39
Depth→Rank	42.24	44.42	58.19
MGCA	**46.60**	**47.86**	**63.15**

Table 2: Downstream accuracy under different depth-derived supervision settings. The three capability groups contain 15,101, 20,337, and 4,635 instances, respectively.

Model	AbsRel ↓	Pair Acc. ↑	Triplet Acc. ↑
Base	0.78	0.10	0.03
Depth-only	0.27	0.84	0.70
Rank-only	22.39	**0.97**	**0.91**
Depth→Rank	0.30	0.95	0.86
MGCA	**0.23**	0.96	0.90

Table 3: Proxy-task performance on the DIODE validation split. AbsRel evaluates metric depth accuracy; Pair Acc. and Triplet Acc. evaluate relative depth ordering.

w_d	w_r	Local Geo.	Relation	Global Org.
0.0	1.0	41.22	43.94	54.39
0.1	0.9	43.96	45.72	57.39
0.2	0.8	29.27	36.15	51.67
0.3	0.7	**46.60**	**47.86**	**63.15**
0.4	0.6	44.11	46.14	<u>62.11</u>
0.5	0.5	43.96	44.29	59.44
0.6	0.4	<u>45.26</u>	<u>46.73</u>	60.00
0.7	0.3	43.80	46.34	58.96
0.8	0.2	39.20	42.74	54.78
0.9	0.1	24.41	32.27	37.73
1.0	0.0	43.97	46.01	59.78

Table 4: Downstream accuracy under different reward compositions. w_d and w_r denote the weights of the metric depth reward and the margin-aware ranking reward, respectively. Best and second-best results are highlighted in bold and underlined.

et al., 2025). These benchmarks provide complementary coverage of structured geometric judgments, relative spatial relations, 3D reasoning in natural imagery, and layout-level spatial understanding. For controlled analysis, we additionally evaluate proxy-level geometric behavior on a fixed three-region set constructed from the DIODE validation split.

4.2 Main Results

Table 1 compares MGCA with representative open-weight and proprietary MLLMs. MGCA is applied to Qwen3-VL-Instruct and requires no additional depth input at inference time.

MGCA improves its backbone on all three benchmarks, with gains of 0.83 points on 3DSRBench, 2.50 points on SpatialEval, and 4.10 points on Spatial457. Among the compared open-weight models of similar scale, MGCA achieves the strongest performance. Together with its consistent gains over the backbone, this result shows that spatial improvements can be obtained through geometry-guided post-training without increasing model size.

Compared with proprietary models, MGCA exhibits a complementary strength. GPT-5 mini and Gemini 3.1 Flash-Lite remain substantially stronger on 3DSRBench and SpatialEval. MGCA, however, obtains the highest score on Spatial457, exceeding Gemini 3.1 Flash-Lite by 3.15 points and GPT-5 mini by 7.87 points. This result suggests that MGCA is particularly effective on the structured spatial judgments represented by Spatial457, even when applied to a compact backbone.

4.3 Understanding Geometric Supervision

We next investigate why MGCA improves downstream spatial understanding. Specifically, we examine how metric and relational supervision affect spatial transfer, and whether performance on the depth-derived proxy task is predictive of downstream spatial reasoning ability.

Downstream transfer under different supervision settings. We first compare MGCA with controlled variants that use the same three-region prediction task but differ in how the geometric supervision is applied. **Depth-only** retains only the metric depth reward, **Rank-only** retains only the margin-aware ranking reward, and **Depth→Rank** applies the two signals sequentially, with depth-based training followed by ranking-based training.

For analysis, we regroup the original benchmark dimensions into three capability categories: *local geometry perception*, *relative spatial relation modeling*, and *global spatial organization and reasoning*. This grouping is used only for analysis and does not alter the original evaluation protocols. We report micro-averaged accuracy over all instances assigned to each category; the detailed mapping is provided in Appendix C.

As shown in Table 2, MGCA is the only setting that improves consistently over the base model across all three capability groups. The improvement is most evident in local geometry perception and relative spatial relation modeling, while the gain on global spatial organization suggests that the learned geometric signal also transfers beyond local depth-related judgments. By contrast,

Variant	3DSRBench	SpatialEval	Spatial457
Base	56.25	60.65	37.91
Supervision strategy			
Single-point SFT	56.38	59.07	37.73
Single-point GRPO	56.53	59.78	38.73
Depth-only	55.29	58.77	39.91
Reward formulation			
Binary ranking reward	55.73	57.61	37.86
Without reward normalization	56.73	57.32	37.30
MGCA	**57.08**	**63.15**	**42.01**

Table 5: Ablation of supervision strategy and reward formulation. Depth-only denotes the three-point variant trained only with the metric depth reward. Binary ranking reward replaces the proposed ranking reward with binary ordering feedback, while Without reward normalization removes online reward normalization.

the controlled variants yield less consistent downstream behavior. Depth-only brings moderate gains on local geometry and relative relations, but does not improve global organization. Rank-only and Depth→Rank fail to match the base model across the three groups, despite directly involving relative-order supervision. This contrast suggests that the downstream benefit of MGCA does not arise from depth prediction or ranking supervision alone.

Proxy-task diagnosis. To further understand these downstream differences, we evaluate the same variants on a fixed three-region proxy set constructed from the DIODE validation split. The proxy evaluation separately measures metric depth accuracy and relative depth-ordering accuracy, allowing us to examine whether success on order prediction is accompanied by metrically grounded predictions. All variants are evaluated on the same proxy instances. Detailed metric definitions are provided in Appendix A.3.

The behavior of Rank-only is particularly revealing. Although it achieves the highest pairwise and triplet ordering accuracy, its metric depth error becomes extremely large and its downstream performance remains poor. This result suggests that relative-order supervision can be optimized successfully without producing metrically grounded predictions.

MGCA reaches a different operating point: it obtains the lowest AbsRel while maintaining ordering accuracy close to Rank-only, and it also performs best across all downstream capability groups. Depth→Rank likewise achieves high proxy-level ordering accuracy, but does not transfer as effectively as joint optimization. Taken together, the proxy and downstream results show that relative ordering becomes more useful for spatial transfer when it is anchored by metric depth supervision.

Reward composition. We further study the balance between the metric depth reward and the ranking reward. As shown in Table 4, $(w_d, w_r) = (0.3, 0.7)$ achieves the highest accuracy in all three capability groups. Neither depth-only nor rank-only matches this joint setting, indicating that both metric and relational supervision contribute to downstream spatial transfer.

Performance also varies substantially across intermediate weight choices. A larger ranking weight is beneficial at the selected operating point, but increasing it further does not consistently improve performance; the same is true for depth-dominant configurations. We therefore use $(0.3, 0.7)$ as the default reward composition, which retains metric grounding while placing greater emphasis on cross-region geometric consistency.

Supervision and reward design. We further ablate the supervision construction and reward formulation in Table 5. **Single-point SFT** trains the model to predict the depth of one sampled region using supervised fine-tuning, while **Single-point GRPO** applies reward-based training to the same single-region task. For the reward design, **Binary ranking reward** replaces the proposed margin-aware ranking reward with binary ordering feedback, and **Without component normalization** removes the online normalization applied to the metric and ranking reward components before aggregation. Implementation details of the reward variants are provided in Appendix B.4.

The single-point variants remain close to the base model, suggesting that learning isolated local depths provides limited benefit for downstream spatial reasoning. Depth-only, which retains the same

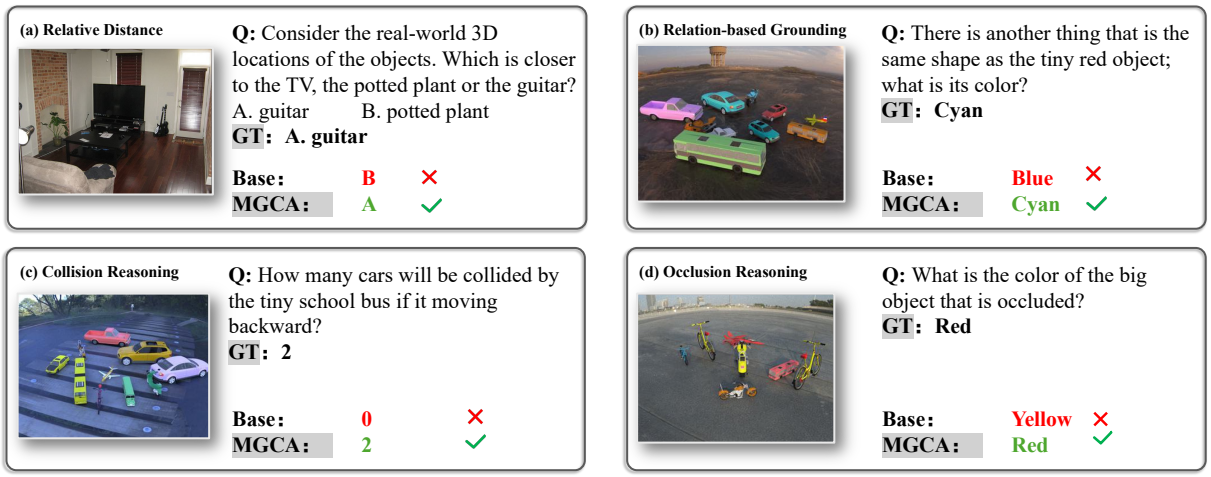


Figure 2: Qualitative comparison between MGCA and its Qwen3-VL-4B-Instruct backbone on downstream spatial reasoning tasks. MGCA corrects errors involving relative distance, relation-based grounding, collision reasoning, and occlusion reasoning. Base denotes the unaligned Qwen3-VL-4B-Instruct model.

three-region construction as MGCA but removes relational supervision, improves Spatial457 but degrades on 3DSRBench and SpatialEval. By contrast, MGCA improves all three benchmarks. This comparison indicates that the benefit of multi-point construction lies not merely in supplying additional depth targets, but in making cross-region geometric relations available for optimization.

The reward ablations lead to the same conclusion. Replacing the margin-aware ranking reward with binary feedback results in clear performance degradation, indicating that coarse ordering supervision does not capture enough geometric structure. Removing component-wise normalization also reduces performance, particularly on SpatialEval and Spatial457, suggesting that balancing the two reward components is important for stable spatial transfer.

4.4 Qualitative Examples

Figure 2 presents representative examples comparing MGCA with its Qwen3-VL-4B-Instruct backbone. The examples cover relative distance, relation-based grounding, collision reasoning, and occlusion reasoning. In each case, the base model fails to resolve the spatial relation required by the question, whereas MGCA produces the correct answer. These examples illustrate that the geometric supervision learned during post-training transfers beyond the proxy depth task to diverse downstream spatial judgments.

5 Conclusion

We present **Multi-point Geometric Consistency Alignment** (MGCA), a geometry-guided post-training method that converts depth annotations into verifiable feedback for improving spatial reasoning in MLLMs. MGCA jointly optimizes metric depth grounding and relative geometric consistency over multiple local regions, without requiring additional depth input at inference time. Experiments demonstrate consistent downstream improvements and show that strong proxy-level ranking accuracy alone does not ensure well-grounded spatial reasoning. By anchoring relational supervision in metric depth feedback, MGCA provides an effective way to strengthen spatial reasoning while preserving the original inference interface.

6 Limitations

MGCA is currently trained with depth supervision from a single dataset and applied to one backbone model. Its effectiveness across other depth domains and MLLM families remains to be evaluated. Moreover, depth-derived supervision mainly improves geometry-related reasoning and does not directly resolve errors arising from object grounding or ambiguous visual evidence. MGCA incurs additional cost during post-training, but requires no extra depth input or computation at inference time.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62576347).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. 2025. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*.

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

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. 2026. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. *Advances in Neural Information Processing Systems*, 38:33938–33968.

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

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Derek Hoiem, Alexei A Efros, and Martial Hebert. 2011. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Disheng Liu, Tuo Liang, Zhe Hu, Jierui Peng, Yiren Lu, Yi Xu, Yun Fu, and Yu Yin. 2025. Spatial intelligence in vision-language models: A comprehensive survey.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. 2008. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan

Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hado P Van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. 2016. Learning values across many orders of magnitude. *Advances in neural information processing systems*, 29.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, and 1 others. 2019. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. 2025b. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24669–24679.

Philip Welsby and Bernard MY Cheung. 2023. Chatgpt.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. 2025b. Cambrian-s: Towards spatial supersensing in video. *Preprint*, arXiv:2511.04670.

Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. 2025. Shapellm-omni: A native multimodal llm for 3d generation and understanding. *arXiv preprint arXiv:2506.01853*.

Hanxun Yu, Xuan Qu, Yuxin Wang, Jianke Zhu, and 1 others. 2026. Unlocking dense metric depth estimation in vlms. *arXiv preprint arXiv:2605.15876*.

Ziyao Zeng, Yangchao Wu, Hyoungseob Park, Daniel Wang, Fengyu Yang, Stefano Soatto, Dong Lao, Byung-Woo Hong, and Alex Wong. 2024. Rsa: Resolving scale ambiguities in monocular depth estimators through language descriptions. *Advances in neural information processing systems*, 37:112684–112705.

Duo Zheng, Shijia Huang, and Liwei Wang. 2025. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8995–9006.

Duo Zheng, Yanyang Li, Liwei Wang, and 1 others. 2026. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586.

Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. 2015. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE international conference on computer vision*, pages 388–396.

A Implementation Details

A.1 Multi-point Training Instance Construction

We construct all geometry-guided post-training instances from the training split of DIODE (Vasiljevic et al., 2019). The split contains 25,458 RGB-depth image pairs, including both indoor and outdoor scenes with substantially different metric depth ranges. As shown in Table 6, outdoor scenes account for a larger portion of the training data and exhibit considerably larger depth ranges than indoor scenes. Using the sampling procedure described below, we obtain 15,575 valid three-region supervision instances, corresponding to an overall construction success rate of 61.18%. The DIODE validation split is excluded from post-training and is used only for the proxy-task analysis in Section 4.3.

Candidate patch filtering. For each depth-annotated image, we sample square local patches of size 8×8. Let p_i denote a candidate patch and Ω_i the set of pixels with valid depth values in p_i. The reference depth of the patch is computed as the mean depth over its valid pixels:

$$g_i = \frac{1}{|\Omega_i|} \sum_{u \in \Omega_i} D(u). \tag{10}$$

Scene	# Images	Median Depth (m)	P90 Depth (m)	# Instances
Indoor	8,574	2.48	6.11	5,068
Outdoor	16,884	8.51	32.69	10,507
Overall	25,458	5.21	26.33	15,575

Table 6: Scene coverage, depth statistics, and constructed multi-point instances from the DIODE training split. The overall construction success rate is 61.18%.

Depth Ratio	Median	Mean
$g_{(2)}/g_{(1)}$	1.71	2.24
$g_{(3)}/g_{(2)}$	1.89	2.69
$g_{(3)}/g_{(1)}$	3.83	5.92

Table 7: Depth-separation statistics of the constructed three-region supervision instances.

A candidate patch is retained only if at least 90% of its pixels have valid depth values and its internal depth variation is limited:

$$\frac{|\Omega_i|}{|p_i|} \geq 0.9, \qquad \frac{\max_{u\in\Omega_i} D(u)}{\min_{u\in\Omega_i} D(u)} \leq 1.15. \quad (11)$$

These constraints remove patches dominated by missing depth values or spanning strong depth discontinuities, for which a single region-level depth target would be unreliable.

Depth-separated triplet sampling. From the retained candidates, we sample three non-overlapping patches and sort their reference depths as $g_{(1)} < g_{(2)} < g_{(3)}$. A triplet is accepted only when adjacent depth levels satisfy

$$\frac{g_{(2)}}{g_{(1)}} \geq 1.5, \qquad \frac{g_{(3)}}{g_{(2)}} \geq 1.5. \quad (12)$$

The selected patches therefore correspond to shallow, middle, and deep local regions within the same image. Table 7 reports the resulting depth-separation statistics. The median depth ratio between the deepest and shallowest regions is 3.83, indicating that the constructed instances contain clear within-image geometric contrasts rather than near-tie comparisons.

Region representation and instruction construction. For each accepted triplet, the three local regions are associated with identifiers A, B, and C. Their locations are represented by bounding boxes normalized to the range $[0, 1000]$ relative to the image width and height:

$$b_i = \left[\frac{1000x_1^{(i)}}{W}, \frac{1000y_1^{(i)}}{H}, \frac{1000x_2^{(i)}}{W}, \frac{1000y_2^{(i)}}{H}\right]. \quad (13)$$

The model receives the original RGB image together with an instruction specifying these three regions by their normalized coordinates. It is asked to predict one metric depth value for each region in a fixed textual format:

```
A: <depth>, B: <depth>, C: <depth>
```

Only responses from which all three depth values can be parsed are assigned non-zero geometric rewards. The colored overlays shown in Figure 1 are included only for visualization and are not part of the model input.

Configuration	Value
Backbone checkpoint	`Qwen/Qwen3-VL-4B-Instruct`
Post-training algorithm	GRPO
Trainable parameters	LoRA adapters
LoRA rank	128
LoRA alpha	256
LoRA dropout	0.05
batch size	128
Number of generations	8
Reward weights (w_d, w_r)	$(0.3, 0.7)$
Epoch	5

Table 8: Training configuration of the main MGCA model.

A.2 Training Configuration

We initialize MGCA from the publicly released `Qwen/Qwen3-VL-4B-Instruct` checkpoint. Post-training is performed with GRPO using LoRA for parameter-efficient optimization. Only the LoRA parameters are updated during post-training. All post-training experiments are conducted on eight NVIDIA RTX A6000 GPUs.

For the main MGCA model, we use the joint geometric consistency reward with $w_d = 0.3$ and $w_r = 0.7$. This configuration assigns greater weight to the ranking component while retaining metric depth supervision for geometric grounding. Unless otherwise specified, the reported MGCA results are obtained with this reward configuration.

Table 8 summarizes the training hyperparameters used for the main model.

A.3 Proxy-task Evaluation

We construct a fixed three-region proxy evaluation set from the DIODE validation split to di-

agnose the geometric behavior learned by different post-training variants. Each proxy instance contains three local regions with reference depths $\mathbf{g} = (g_1, g_2, g_3)$, and the evaluated model produces the corresponding predictions $\hat{\mathbf{d}} = (\hat{d}_1, \hat{d}_2, \hat{d}_3)$. All variants are evaluated on the same set of proxy instances.

The proxy evaluation reports metric depth accuracy and relative depth-ordering accuracy separately. These metrics are used only for analysis and are distinct from the reward functions used during post-training. In particular, the metric evaluation uses standard absolute relative error rather than the symmetric relative term in the training reward, while the ordering metrics assess discrete correctness rather than the margin-aware ranking score.

Absolute relative error. Let $\mathcal{V}$ denote the set of valid predicted region points whose reference and predicted depths are finite and whose reference depth is positive. We compute Absolute Relative Error (AbsRel) as

$$\mathrm{AbsRel} = \frac{1}{|\mathcal{V}|} \sum_{(n,i)\in\mathcal{V}} \frac{\left|\hat{d}_i^{(n)} - g_i^{(n)}\right|}{g_i^{(n)}}, \tag{14}$$

where n indexes proxy instances and $i \in \{1, 2, 3\}$ indexes the three sampled regions. Lower AbsRel indicates more accurate metric depth grounding.

Pairwise rank accuracy. For each valid triplet n, we consider the three region pairs

$$\mathcal{E} = \{(1, 2), (1, 3), (2, 3)\}. \tag{15}$$

Pairs whose reference depth difference is within a numerical tie tolerance are excluded from evaluation. Let

$$\mathcal{P} = \left\{(n, i, j) \;\middle|\; (i, j) \in \mathcal{E}, \left|g_i^{(n)} - g_j^{(n)}\right| > \varepsilon_{\mathrm{eval}}\right\} \tag{16}$$

denote the set of evaluated non-tied pairs. For each evaluated pair, we define the ordering correctness indicator as

$$\begin{aligned} c_{ij}^{(n)} = \mathbf{1}\Big[&\mathrm{sgn}_\varepsilon\big(\hat{d}_i^{(n)} - \hat{d}_j^{(n)}\big) \\ &= \mathrm{sgn}_\varepsilon\big(g_i^{(n)} - g_j^{(n)}\big)\Big], \end{aligned} \tag{17}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function and $\mathrm{sgn}_\varepsilon(\cdot)$ returns zero for values whose absolute magnitude is no larger than $\varepsilon_{\mathrm{eval}}$. Pairwise rank accuracy is then computed as

$$\mathrm{PairAcc} = \frac{1}{|\mathcal{P}|} \sum_{(n,i,j)\in\mathcal{P}} c_{ij}^{(n)}. \tag{18}$$

In implementation, we set $\varepsilon_{\mathrm{eval}} = 10^{-8}$.

Triplet order accuracy. Triplet order accuracy evaluates whether the predicted depths recover the complete ordering of the three regions. For each valid triplet $n \in \mathcal{T}$, we define

$$c_{\mathrm{tri}}^{(n)} = \mathbf{1}\left[\mathrm{Order}\big(\hat{\mathbf{d}}^{(n)}\big) = \mathrm{Order}\big(\mathbf{g}^{(n)}\big)\right], \tag{19}$$

where $\mathrm{Order}(\cdot)$ returns the ascending ordering of the three regions according to their depth values. Triplet order accuracy is then computed as

$$\mathrm{TripletAcc} = \frac{1}{|\mathcal{T}|} \sum_{n\in\mathcal{T}} c_{\mathrm{tri}}^{(n)}. \tag{20}$$

B Details of the Geometric Consistency Reward

The geometric consistency reward evaluates each generated response from two complementary perspectives: metric depth accuracy at individual regions and relative geometric consistency across regions. For a parsed response $\hat{\mathbf{d}} = (\hat{d}_1, \hat{d}_2, \hat{d}_3)$ and its reference depths $\mathbf{g} = (g_1, g_2, g_3)$, the overall reward is defined as

$$R = r_{\mathrm{fmt}} \left[w_d \widetilde{R}_{\mathrm{depth}} + w_r \widetilde{R}_{\mathrm{rank}}\right], \tag{21}$$

where $r_{\mathrm{fmt}} \in \{0, 1\}$ indicates whether all three predicted depths can be successfully parsed from the required output format. For the main MGCA model, we set $(w_d, w_r) = (0.3, 0.7)$.

B.1 Metric Depth Reward

The metric component evaluates the numerical plausibility of the predicted depth for each sampled region. Since DIODE contains both indoor and outdoor scenes with substantially different depth ranges, we combine a scale-aware relative error with a clipped absolute-error term.

For region i, the symmetric relative error is computed as

$$e_i^{\mathrm{rel}} = \frac{2|\hat{d}_i - g_i|}{|\hat{d}_i| + |g_i| + \varepsilon}, \tag{22}$$

where ε is a small constant for numerical stability. This term measures prediction error relative to the depth scale of the corresponding region.

We additionally compute a clipped absolute hinge error:

$$e_i^{\mathrm{abs}} = \mathrm{clip}\left(|\hat{d}_i - g_i| - \tau_{\mathrm{abs}}, 0, c_{\mathrm{abs}}\right), \quad (23)$$

where τ_{abs} specifies a tolerance margin and c_{abs} limits the influence of extremely large absolute deviations. The point-wise depth reward is then defined as

$$r_i^{\mathrm{depth}} = \exp\left(-\alpha_{\mathrm{rel}} e_i^{\mathrm{rel}} - \beta_{\mathrm{abs}} e_i^{\mathrm{abs}}\right), \quad (24)$$

and the metric depth reward for a triplet is obtained by averaging over its three regions:

$$R_{\mathrm{depth}} = \frac{1}{3}\sum_{i=1}^{3} r_i^{\mathrm{depth}}. \quad (25)$$

In our implementation, we set $\alpha_{\mathrm{rel}} = 2.0$, $\beta_{\mathrm{abs}} = 2.0$, $\tau_{\mathrm{abs}} = 0.01$, and $c_{\mathrm{abs}} = 5.0$. The relative term prevents large-range outdoor scenes from dominating the reward solely because of their larger metric depths, while the clipped absolute term discourages predictions with large physical deviations.

B.2 Margin-aware Ranking Reward

The metric depth reward evaluates each sampled region independently. To further measure whether the predicted depths preserve the cross-region geometry of a triplet, we define a margin-aware ranking reward over all region pairs:

$$\mathcal{E} = \{(1,2),(1,3),(2,3)\}. \quad (26)$$

For each pair $(i,j) \in \mathcal{E}$, we compute the predicted and reference depth margins:

$$\Delta_{ij} = \hat{d}_i - \hat{d}_j, \qquad \Delta_{ij}^* = g_i - g_j. \quad (27)$$

Margin clipping and pair weighting. Large metric depth gaps may otherwise dominate the ranking reward, particularly across scenes with different depth ranges. We therefore clip both the predicted and reference margins to a bounded interval:

$$\begin{aligned} \bar{\Delta}_{ij} &= \mathrm{clip}\left(\Delta_{ij}, -c_m, c_m\right), \\ \bar{\Delta}_{ij}^* &= \mathrm{clip}\left(\Delta_{ij}^*, -c_m, c_m\right). \end{aligned} \quad (28)$$

We also reduce the influence of near-tie reference pairs by assigning each pair a weight

$$\omega_{ij} = \min\left(\frac{|\Delta_{ij}^*|}{\tau_{\mathrm{tie}}}, 1\right). \quad (29)$$

Although the triplet construction already favors depth-separated regions, this weighting provides additional robustness when a pair has a small reference margin.

Order-consistency score. The order-consistency term evaluates whether the predicted margin has the same near–far direction as the reference margin:

$$s_{ij}^{\mathrm{ord}} = \sigma\left(k \frac{\bar{\Delta}_{ij}^*}{|\Delta_{ij}^*|} \bar{\Delta}_{ij}\right), \quad (30)$$

where $\sigma(\cdot)$ is the sigmoid function and k controls the sharpness of the ordering score. For a zero-gap reference pair, the order-consistency score is set to 0.5.

Gap-consistency score. Correct ordering alone does not distinguish between metrically plausible predictions and predictions that preserve only a coarse relative order. We therefore additionally measure whether the magnitude of the predicted pairwise margin agrees with that of the reference margin:

$$s_{ij}^{\mathrm{gap}} = \exp\left(-\frac{\left||\bar{\Delta}_{ij}| - |\bar{\Delta}_{ij}^*|\right|}{\tau_{\mathrm{gap}}}\right), \quad (31)$$

where τ_{gap} controls the tolerance for margin mismatch.

We aggregate the order-consistency and gap-consistency scores over the three region pairs:

$$R_q = \frac{\sum_{(i,j)\in\mathcal{E}} \omega_{ij} s_{ij}^q}{\sum_{(i,j)\in\mathcal{E}} \omega_{ij}}, \qquad q \in \{\mathrm{ord}, \mathrm{gap}\}. \quad (32)$$

The final ranking reward is defined as

$$R_{\mathrm{rank}} = \lambda R_{\mathrm{ord}} + (1-\lambda) R_{\mathrm{gap}}. \quad (33)$$

In our implementation, we set $k = 2.0$, $\tau_{\mathrm{gap}} = 0.05$, $\tau_{\mathrm{tie}} = 0.15$, $c_m = 5.0$, and $\lambda = 0.6$.

B.3 Component-wise Online Normalization

The metric depth reward and the ranking reward evaluate different aspects of geometric consistency and may exhibit different value distributions during training. Directly combining the two raw components can therefore lead to an imbalanced optimization signal. To alleviate this issue, we maintain separate online statistics for R_{depth} and R_{rank}.

For each reward component $c \in \{\mathrm{depth}, \mathrm{rank}\}$, we compute

$$\begin{aligned} z_c &= \mathrm{clip}\left(\frac{R_c - \mu_c}{\sqrt{\sigma_c^2 + \varepsilon}}, -\gamma, \gamma\right), \\ \widetilde{R}_c &= \sigma(z_c), \end{aligned} \quad (34)$$

where μ_c and σ_c^2 denote the running mean and variance of component c, respectively. The running statistics are updated using exponential moving averages of batch-level statistics, which are synchronized across devices during distributed training. Normalization is activated after a warm-up stage; before that, the raw reward components are used directly.

In our implementation, we set the exponential moving average coefficient to 0.98, the clipping threshold to $\gamma = 3.0$, and the warm-up duration to 200 optimization steps. This normalization is applied to the two reward components before their weighted aggregation in Eq. 21, and is separate from the group-relative processing performed by GRPO.

B.4 Reward Variants in Ablation

We construct two reward variants to examine the effects of the margin-aware ranking formulation and component-wise normalization.

Binary ranking reward. This variant retains the three-region supervision construction and the metric depth component of MGCA, but replaces the margin-aware ranking reward with binary pairwise ordering feedback. For each pair $(i, j) \in \mathcal{E}$, we define

$$s_{ij}^{\text{bin}} = \mathbf{1}\left[\Delta_{ij}\Delta_{ij}^{*} > 0\right], \tag{35}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, and $\Delta_{ij} = \hat{d}_i - \hat{d}_j$ and $\Delta_{ij}^{*} = g_i - g_j$ are the predicted and reference depth margins. The binary ranking reward is computed as

$$R_{\text{bin}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} s_{ij}^{\text{bin}}. \tag{36}$$

Since the triplet construction enforces clearly separated reference depths, the evaluated pairs do not contain ambiguous near-ties. Unlike the proposed margin-aware ranking reward, this variant evaluates only order correctness and provides no feedback on the magnitude of pairwise depth margins.

Without component normalization. This variant removes the component-wise online normalization described in Appendix B.3. The metric depth reward and the margin-aware ranking reward are combined directly in their original scales:

$$R = r_{\text{fmt}}\left[w_d R_{\text{depth}} + w_r R_{\text{rank}}\right]. \tag{37}$$

All other components of the reward formulation remain unchanged.

C Capability Grouping for Downstream Analysis

To analyze how geometric supervision transfers to different forms of spatial reasoning, we regroup the original evaluation dimensions of Spatial457, SpatialEval, and 3DSRBench into three capability groups: *local geometry perception*, *relative spatial relation modeling*, and *global spatial organization and reasoning*. This regrouping is used only for capability-level analysis and does not modify the original evaluation protocol or benchmark instances. For each group, we report micro-averaged accuracy over all instances belonging to the assigned original dimensions.

Local geometry perception. This group focuses on geometric judgments involving individual objects or local interactions, including local scale, depth ordering, occlusion, collision, height, and camera-relative distance. It contains the object-level, occlusion, and collision dimensions from Spatial457, together with height and camera-distance dimensions from 3DSRBench.

Relative spatial relation modeling. This group evaluates whether the model can reason about spatial relations among multiple objects, including relative position, orientation, viewpoint, facing direction, parallelism, and multi-object distance comparison. It contains the relational and pose-related dimensions from Spatial457 and the corresponding relation-oriented dimensions from 3DSRBench.

Global spatial organization and reasoning. This group measures spatial reasoning over larger layouts or structured environments, including maps, grids, maze navigation, and real-scene layout understanding. It contains all four task dimensions of SpatialEval.

D Additional Evaluation and Robustness Analysis

This section reports the additional experiments conducted in response to reviewer feedback. We focus on the statistical reliability of the downstream gains, possible regressions in general visual–language capability, transfer to a second model family, and sensitivity to two design factors: patch size and depth quality. Unless otherwise stated, results in this section use the Qwen3-VL-4B backbone; Section D.3 instead evaluates InternVL3.5-4B.

Capability Group	Benchmark	Included Original Dimensions	Group Size
Local geometry perception	Spatial457	`L1_single`, `L2_objects`, `L4_occ`, `L5_collision`	15,101
	3DSRBench	`height_higher`, `location_closer_to_camera`	–
Relative spatial relation modeling	Spatial457	`L3_2d_spatial`, `L4_pose`, `L5_6d_spatial`	20,337
	3DSRBench	`location_above`, `location_next_to`, `orientation_on_the_left`, `orientation_in_front_of`, `orientation_viewpoint`, `multi_object_closer_to`, `multi_object_facing`, `multi_object_parallel`, `multi_object_same_direction`, `multi_object_viewpoint_towards_object`	–
Global spatial organization and reasoning	SpatialEval	`spatialmap`, `spatialgrid`, `mazenav`, `spatialreal`	4,635

Table 9: Mapping from original benchmark dimensions to the capability groups used for downstream analysis. Group Size denotes the total number of instances assigned to each capability group. The grouping is used only for analysis, and accuracy is micro-averaged over all assigned instances.

D.1 Statistical Reliability of Downstream Gains

Because the absolute improvement on 3DSRBench is modest, we complement point estimates with paired, question-level statistical tests. Base-model and MGCA predictions are matched on the same evaluation examples. We estimate two-sided 95% paired-bootstrap confidence intervals for the accuracy difference and apply an exact two-sided McNemar test to the discordant prediction pairs. Table 10 reports the resulting intervals and p-values.

All three confidence intervals lie above zero, and the corresponding McNemar tests reject equal error rates at the 0.05 level. These results support the reliability of the improvements under test-set sampling variation. Nevertheless, the 0.83-point gain on 3DSRBench remains small in magnitude and should not be interpreted as closing the gap to substantially larger proprietary models.

D.2 General Visual–Language Capability

We additionally evaluate the base model and MGCA on MMMU-Pro and MMBench to examine whether geometry-guided post-training causes a broad regression in non-spatial capability. Table 11 shows that the two text-option settings of MMMU-Pro improve by 1.39 and 1.73 points, whereas MMMU-Pro Vision and MMBench validation decrease by 1.10 and 1.04 points, respectively.

The changes are small relative to the improvements on Spatial457 and SpatialEval, suggesting that MGCA does not cause a broad collapse of general visual–language capability. At the same time, the decreases on two settings show that post-training does not uniformly preserve every capability; we therefore treat general-capability retention as an empirical trade-off rather than an unconditional guarantee.

D.3 Cross-backbone Transfer to InternVL3.5-4B

To examine whether MGCA is tied to the Qwen3-VL architecture, we apply the same geometry-guided post-training formulation to InternVL3.5-4B (Wang et al., 2025a). The base and adapted checkpoints are evaluated under the same InternVL inference configuration. As shown in Table 12, MGCA improves 3DSRBench by 1.62 points and SpatialEval by 4.57 points, while Spatial457 decreases by 2.69 points.

These results provide evidence that the training formulation can transfer to a second MLLM family, but they do not establish architecture-independent improvement. The InternVL evaluation uses a constrained visual-token budget for resource reasons. Because Spatial457 emphasizes fine-grained multi-object binding, occlusion, pose, and 6D relations, its decline may reflect sensitivity to the visual representation and inference configuration. The present experiment does not isolate that cause, however, so we report the mixed outcome as a limitation rather than attributing the decrease solely to token resolution.

D.4 Patch-size Sensitivity

The default 8×8 patches are intended to retain locally coherent geometry and reduce the probability that a sampled region crosses an object boundary or depth discontinuity. We test a larger 16×16 patch using the checkpoint after 2,000 optimization steps,

Benchmark	Base	MGCA	Gain	95% Paired-bootstrap CI	McNemar p
3DSRBench	56.25	57.08	$+0.83$	$[+0.12, +1.54]$	0.023
Spatial457	37.91	42.01	$+4.10$	$[+3.57, +4.64]$	2.38×10^{-51}
SpatialEval	60.65	63.15	$+2.50$	$[+1.53, +3.47]$	5.59×10^{-7}

Table 10: Question-level statistical analysis of MGCA over its Qwen3-VL-4B backbone. Scores and gains are in percentage points. All paired-bootstrap intervals exclude zero; McNemar tests are exact and two-sided.

Evaluation	Base	MGCA	Diff.
MMMU-Pro std. (10 options)	33.01	34.39	$+1.39$
MMMU-Pro std. (4 options)	44.86	46.59	$+1.73$
MMMU-Pro Vision	25.66	24.57	-1.10
MMBench validation	89.00	87.96	-1.04

Table 11: General visual–language benchmark results. "Diff." denotes MGCA minus the base model.

Model / setting	3DSR	SEval	S457
InternVL3.5-4B	52.62	48.89	34.30
InternVL3.5-4B + MGCA	54.24	53.46	31.61
Difference	$+1.62$	$+4.57$	-2.69

Table 12: Cross-backbone evaluation on InternVL3.5-4B. "Difference" denotes MGCA minus the base checkpoint under the same InternVL evaluation configuration.

while keeping the remaining training and evaluation settings unchanged. As shown in Table 13, the larger patch is worse on all three downstream benchmarks. This result favors compact local regions under the present construction, although the comparison is a two-point sensitivity check rather than an exhaustive sweep over region scales.

D.5 Sensitivity to Depth Quality

To test whether MGCA actually depends on the geometric supervision, we perturb the training-time depth annotations with Gaussian noise while leaving the remaining training procedure unchanged. Table 14 shows consistent degradation under noisy supervision, with the largest drops on Spatial457 and SpatialEval. The result indicates that MGCA uses the depth-derived feedback rather than improving through an unrelated training shortcut, while also exposing depth quality as an important boundary condition. Valid-depth masking, local-stability filtering, depth-separation constraints, reward clipping, and online normalization reduce the influence of ambiguous regions, but they do not make the method invariant to corrupted supervision.

Patch size	3DSRBench	Spatial457	SpatialEval
8×8	56.82	41.79	62.89
16×16	55.50	39.15	60.37

Table 13: Patch-size sensitivity for Qwen3-VL-4B + MGCA at optimization step 2,000.

Training signal	3DSR	S457	SEval
Clean depth	56.82	41.79	62.89
Gaussian-noise depth	56.21	35.77	55.02

Table 14: Sensitivity of Qwen3-VL-4B + MGCA to Gaussian perturbations of the training-time depth annotations.

E Additional Comparisons and Scope

E.1 Comparison with Closely Related Spatial Models

Table 15 evaluates public SpatialRGPT and SpatialThinker checkpoints under our benchmark protocol. The comparison is informative but not a controlled ablation: the models differ in parameter count, training data, supervision, and inference interface. In particular, SpatialRGPT uses region-grounded spatial representations, SpatialThinker performs structured spatial reasoning, whereas MGCA preserves the original RGB–text interface of its Qwen3-VL backbone.

SpatialVLM (Chen et al., 2024) is included in our methodological discussion but not in the numerical table because we did not identify a public checkpoint or fully reproducible evaluation configuration suitable for the same protocol. SpatialVLM primarily constructs quantitative spatial instruction data for supervised training, while MGCA turns metric depth annotations into rule-computable rewards for GRPO-based post-training. Thus, the contribution of MGCA is not depth supervision, ranking, or GRPO in isolation, but their combination in a training-time-only formulation that jointly aligns metric grounding and relative geometric consistency.

Method	Inference interface	# Params	3DSRBench	Spatial457	SpatialEval
SpatialRGPT (Cheng et al., 2024)	Region/depth-aware	8B	51.90	36.27	38.32
SpatialThinker (Batra et al., 2025)	Scene-graph reasoning	3B	44.36	26.66	29.43
Qwen3-VL-4B + MGCA	RGB–text	4B	**57.08**	**42.01**	**63.15**

Table 15: Horizontal comparison with closely related open-weight spatial-reasoning models under our evaluation protocol. SpatialRGPT uses the VILA1.5-8B checkpoint. Differences in architecture, training data, and inference interface preclude a controlled method-only comparison.

E.2 Depth as Feedback versus Depth as Input

Depth-as-input and depth-as-feedback methods address different deployment assumptions. The former enriches inference with an additional geometric observation, but its behavior depends on the depth estimator, the encoding and fusion design, and the availability of depth maps or region features at deployment. MGCA instead uses depth only to construct verifiable rewards during post-training. At inference time, it requires neither a depth estimator nor depth tokens, region proposals, or a scene graph, and therefore retains the standard RGB–text interface. This distinction should not be read as a claim that depth as feedback universally dominates depth as input; the two paradigms can be complementary when depth observations are available.

E.3 Scaling beyond Three Regions and Boundary Conditions

Three regions provide a near–middle–far structure and three pairwise constraints while keeping sampling, structured generation, and reward computation manageable. The construction extends directly to K regions, yielding $K(K-1)/2$ pairwise constraints. Increasing K may enrich the supervision, but it also makes valid region sampling harder, lengthens the output sequence, increases parsing failures and reward variance, and raises training cost. A systematic study of this trade-off remains future work.

The current evidence is further bounded by depth supervision from DIODE and two 4B open-weight backbones. Qwen3-VL-4B improves on all three spatial benchmarks, whereas the InternVL3.5-4B transfer result is mixed; consequently, these experiments do not establish architecture-independent gains. MGCA may also fail when local evidence is ambiguous, when a sampled patch crosses a depth discontinuity, or when the downstream error is dominated by object recognition or grounding rather than geometry. Proprietary models are therefore retained only as reference points; our empirical claim is limited to the evaluated backbone–benchmark combinations, not that MGCA closes the gap to substantially larger closed models.

F Artifact Licenses and Intended Use

We use existing datasets, model checkpoints, and benchmarks only for research-oriented post-training and evaluation. DIODE (Vasiljevic et al., 2019), from which we construct the multi-point supervision instances, is released under the MIT License. The Qwen3-VL-4B-Instruct checkpoint used for the main experiments is released under the Apache License 2.0. The public InternVL3.5-4B checkpoint is used only for the cross-backbone evaluation in Section D.3. Among the evaluation benchmarks, Spatial457 (Wang et al., 2025b) is released under the Apache License 2.0, and 3DSR-Bench (Ma et al., 2025) is released under CC BY 4.0. For SpatialEval (Wang et al., 2024), we did not identify an explicit public license statement at the time of writing; we therefore use it only for evaluation and do not redistribute its contents. All derived supervision instances and trained models in this work are intended for research use consistent with the access conditions of the original artifacts.