GeoMem3D: Learning to Reason over Structured 3D World Memory

Anonymous authors
Paper under double-blind review

Abstract

Multimodal Large Language Models (MLLMs) have significantly advanced 3D spatial reasoning, yet their ability to maintain consistent spatial understanding across viewpoints remains limited. Recent methods improve spatial reasoning by extracting evidence from continuous visual observations or interactive exploration. However, this paradigm is constrained by observation-level evidence, limiting consistent spatial reasoning across views and queries. Motivated by this observation, we propose GeoMem3D, a learning framework that enables MLLMs to reason over a persistent and structured 3D world memory of all objects and their spatial relations. GeoMem3D reconstructs multi-view scenes within a shared geometric frame and organizes object entities, geometric attributes, and spatial relations into a query-independent structured 3D world memory, which provides persistent and reusable spatial knowledge for reasoning. Leveraging this memory, we explore two complementary reasoning paradigms: GeoMem-Holistic directly reasons over the complete structured memory, while GeoMem-Agentic adaptively accesses fine-grained spatial evidence from the memory during reasoning. We further introduce a two-stage learning strategy that combines supervised fine-tuning with reinforcement learning to establish effective memory utilization behaviors. Experiments on VSI-Bench-Tiny and MindCube-Tiny demonstrate consistent improvements across diverse spatial reasoning tasks, validating the effectiveness and generalizability of structured 3D world memory for spatial reasoning.

1 Introduction

3D spatial understanding has become a prerequisite and enabling capability for emerging embodied intelligence (Liu et al., 2026b; Yuan et al., 2026; Zhang et al., 2026a). Although multimodal large language models (MLLMs) have advanced open-ended visual understanding and multimodal reasoning (Liu et al., 2023; Bai et al., 2025b; Zhu et al., 2025), their ability to perform reliable 3D spatial reasoning remains limited, particularly in maintaining coherent spatial representations. Achieving reliable 3D spatial reasoning, including metric geometry recovery, spatial relation reasoning, and navigation, requires models to integrate incomplete observations and establish coherent geometric and semantic relationships across views. However, occlusion and viewpoint variation make cross-view correspondence difficult and hinder coherent scene understanding.

To bridge this gap, existing methods improve 3D spatial reasoning from several perspectives. Some methods (Chen et al., 2024; Wu et al., 2025) leverage large-scale geometry-grounded supervision to train MLLMs for spatial reasoning from 2D observations. Although these methods improve spatial reasoning, the acquired spatial knowledge remains implicit in model parameters or latent representations, requiring MLLMs to infer 3D structures from viewpoint-dependent observations. Other approaches (Hong et al., 2023; Fan et al., 2026) couple MLLMs with explicit 3D inputs or reconstruction-derived representations, providing stronger geometric information but requiring specialized encoders and large-scale 3D-language or reconstruction supervision. Moreover, the resulting geometry is typically encoded as latent features or tokens rather than explicit entities and relations, limiting their ability to support explicit and reusable spatial reasoning.

Subsequent methods (Ma et al., 2024; Zhang et al., 2026b) move beyond fixed 3D encoding by allowing MLLMs to invoke geometric tools or manipulate reconstructed scenes during inference. Acquiring such evidence is query-dependent and often requires multiple rounds of model and tool

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

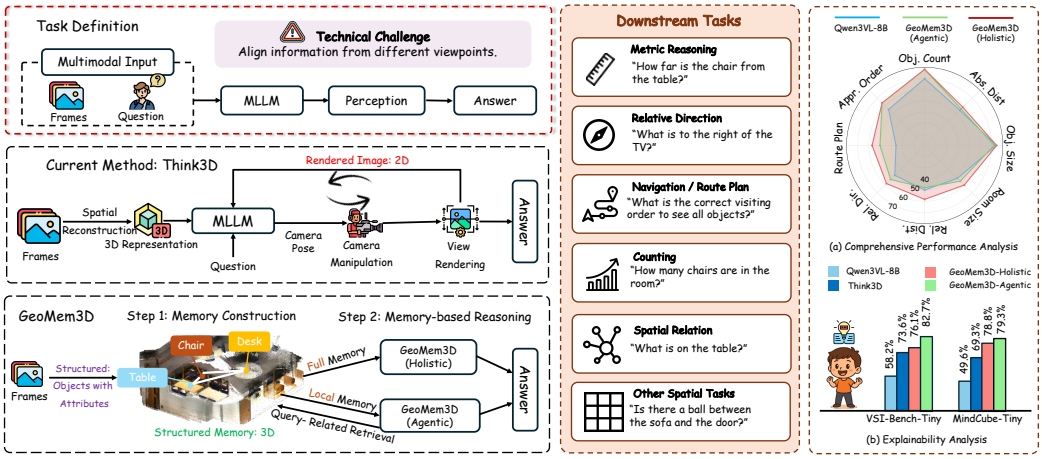


Figure 1: **From View-Dependent Observations to Structured 3D World Memory.** Recent approaches such as Think3D (Zhang et al., 2026b) improve spatial reasoning by acquiring additional 2D visual evidence through interactive viewpoint manipulation. GeoMem3D instead constructs a persistent structured 3D world memory from multi-view observations, enabling MLLMs to perform memory-grounded spatial reasoning through holistic reasoning and agentic memory access.

interaction, introducing additional inference overhead. Fundamentally, viewpoint manipulation typically reprojects the reconstructed geometry into 2D observations, requiring the model to infer spatial relations again from visual evidence, as shown in Figure 1.

These limitations motivate a different perspective: rather than repeatedly acquiring task-specific 3D evidence from observations, MLLMs could reason directly over a persistent and structured representation of 3D entities and their spatial relations, referred to as 3D world memory, which provides a query-independent representation of the environment by organizing object entities, geometric attributes, spatial relations, and other structured spatial information while maintaining cross-view consistency. Therefore, we introduce **GeoMem3D**, a unified framework integrating perception, reconstruction, memory, and reasoning. Specifically, we investigate two complementary modes of reasoning over the same underlying structured 3D world memory, as shown in Figure 1. **GeoMem-Holistic** provides MLLMs with the complete structured memory and original visual context, enabling direct reasoning over the global scene layout, object entities, and spatial relations. **GeoMem-Agentic**, by contrast, adaptively accesses the structured memory according to the query and evolving reasoning state, obtaining fine-grained evidence at the object, relation, region, or measurement level when more specific spatial clues are required. We further train these reasoning modes through a two-stage learning strategy that first establishes memory-grounded reasoning behaviors via supervised fine-tuning and then refines memory utilization through GRPO-based reinforcement learning. In this way, GeoMem3D connects 3D perception, persistent structured memory, and MLLM reasoning, explicitly modeling how structured 3D memory supports spatial reasoning. Experiments on VSI-Bench-Tiny and MindCube-Tiny show that GeoMem3D improves performance by up to 8.6% and 4.6% over Qwen3-VL-8B, demonstrating the effectiveness and generalization of our method.

Our main contributions are summarized as follows:

- We propose **GeoMem3D**, a unified perception, reconstruction, memory, and reasoning framework that introduces a persistent and query-independent structured 3D world memory for MLLM spatial reasoning.
- We introduce two complementary memory-grounded reasoning paradigms over the same structured memory: **GeoMem-Holistic** reasons over the complete memory, while **GeoMem-Agentic** adaptively accesses fine-grained spatial evidence based on the query and reasoning process.
- We develop a two-stage learning strategy that combines supervised fine-tuning and GRPO-based reinforcement learning to establish effective memory utilization behaviors.

- Experiments on VSI-Bench-Tiny and MindCube-Tiny demonstrate consistent improvements across diverse spatial reasoning tasks, validating the effectiveness and generalizability of structured 3D world memory.

2 Related Work

2.1 Multimodal Large Language Model Reasoning

Recent MLLMs (Liu et al., 2023; Bai et al., 2025b; Zhu et al., 2025; Li et al., 2024; Team et al., 2024) have substantially advanced the joint understanding of text, images, and videos by coupling pretrained visual encoders with large language models and scaling multimodal pretraining and instruction tuning. Building on these general-purpose models, some studies (Xu et al., 2025; Huang et al., 2026a; Zhang et al., 2025; Guo et al., 2025; Li et al., 2026b) have strengthened multimodal reasoning through structured reasoning supervision, inference-time scaling, and reinforcement-learning-based post-training. These advances enable more systematic multi-step inference and more effective integration of visual and linguistic evidence across a broad range of tasks (Deshpande et al., 2025; Li et al., 2023). Nevertheless, 3D spatial understanding remains challenging because it requires models to consolidate partial, viewpoint-dependent observations into a coherent geometric representation and infer metric relationships that are not directly available from individual 2D inputs.

2.2 3D Spatial Reasoning

Prior 3D vision research has developed explicit scene representations for geometric perception and rendering, including point-cloud encoding Qi et al. (2017), 3D object detection Qi et al. (2019), relational scene modeling Wald et al. (2020), and 3D Gaussian Splatting Kerbl et al. (2023). However, these methods are typically optimized for specific tasks such as recognition, detection, reconstruction, or rendering, offering limited support for holistic and open-ended 3D scene understanding.

Recent advances in MLLMs have extended multimodal reasoning toward 3D spatial understanding (Hong et al., 2023; Daxberger et al., 2025; Ma et al., 2026; Liu et al., 2026a). A prominent line of work improves spatial reasoning through large-scale geometry-grounded supervision. SpatialVLM (Chen et al., 2024) learns metric spatial priors from generated 3D question-answer data, whereas Spatial-MLLM (Wu et al., 2025) combines semantic and geometry-aware representations with instruction tuning and reinforcement learning. Despite these improvements, spatial knowledge remains implicit in model parameters or latent representations, requiring MLLMs to infer 3D geometry from viewpoint-dependent observations.

Several works (Hong et al., 2023; Fan et al., 2026; Huang et al., 2026b; Wang et al., 2025) incorporate explicit 3D inputs or reconstruction-derived representations into MLLMs for spatial reasoning. 3D-LLM (Hong et al., 2023) projects point-cloud features into a language model for 3D grounding, question answering, and navigation, while VLM-3R (Fan et al., 2026) extracts implicit scene and viewpoint representations from videos through reconstructive instruction tuning. While providing stronger geometric grounding, these approaches require specialized 3D inputs, geometry encoders, or reconstruction supervision, and encode geometry in architecture-specific latent representations, limiting their flexibility for explicit and reusable spatial reasoning.

More recent approaches (Ma et al., 2024; Zhang et al., 2026b; Chen et al., 2026; Zhong et al., 2026; Marsili et al., 2025) explore interactive spatial reasoning by allowing MLLMs to acquire additional spatial evidence through tool invocation and adaptive viewpoint transformations. SpatialPIN (Ma et al., 2024) coordinates complementary priors from multiple 3D foundation models through training-free interaction, while Think3D (Zhang et al., 2026b) reconstructs point clouds and camera poses and allows MLLMs to explore the resulting space through viewpoint manipulation and global–egocentric view switching. This paradigm turns geometry from a static input into an interactive medium for reasoning. However, this paradigm incurs the overhead of iterative tool invocation and repeated rendering. More fundamentally, reconstructed geometry often serves only to generate new 2D observations, leaving MLLMs to re-infer spatial relations from visual evidence rather than reason directly over explicit 3D structure. GeoMem3D instead organizes reconstructed geometry into structured, object-centric memory, such that 3D entities and their relations form the basis of reasoning rather than auxiliary visual evidence.

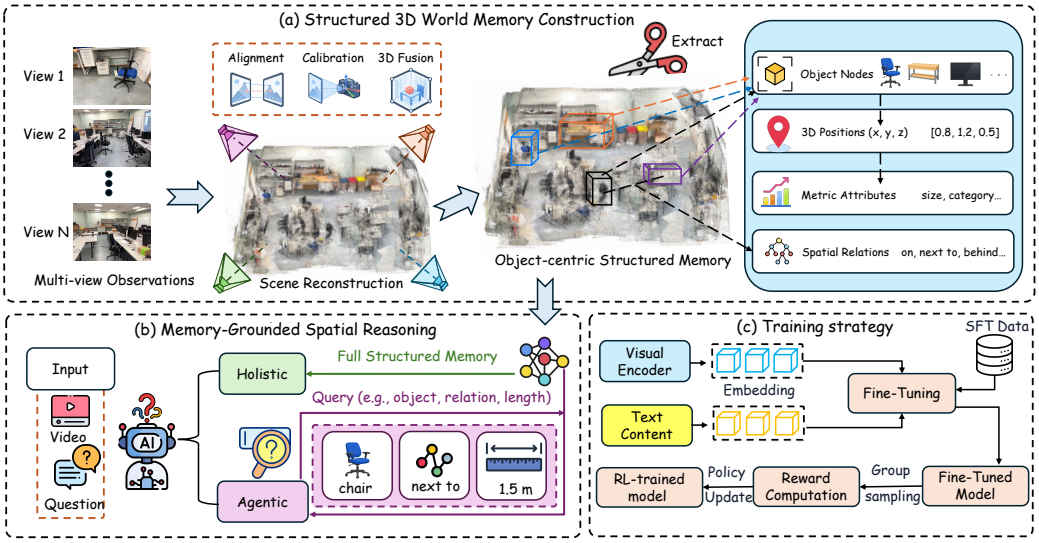


Figure 2: **Overview of Our Proposed Method.** (a) Structured 3D World Memory Construction: multi-view observations are reconstructed and organized into a persistent, query-independent, object-centric structured 3D world memory. (b) Memory-Grounded Spatial Reasoning: MLLMs reason over the structured memory through two complementary paradigms, where GeoMem-Holistic reasons over the complete memory and GeoMem-Agentic accesses query-relevant spatial evidence. (c) Training Strategy: memory-grounded reasoning behaviors are established through supervised fine-tuning and optimized with GRPO-based reinforcement learning.

3 Methodology

3.1 Overview

As illustrated in Figure 2, GeoMem3D formulates 3D spatial reasoning as learning to reason over a persistent and structured 3D world memory. It first reconstructs multi-view observations into a unified 3D scene and organizes the resulting geometry into object entities, geometric attributes, spatial relations and other spatial information, forming an object-centric structured 3D world memory (Section 3.2). This memory supports two complementary reasoning modes: GeoMem-Holistic reasons over the complete structured memory, whereas GeoMem-Agentic adaptively accesses query-relevant spatial evidence from the memory during reasoning (Section 3.3). We train both modes through memory-grounded supervised fine-tuning followed by GRPO-based reinforcement learning to refine effective memory utilization behaviors (Section 3.4).

3.2 Structured 3D World Memory

Given a set of multi-view RGB observations $\mathcal{V} = \{I_i\}_{i=1}^{N}$ of the same physical environment, where I_i denotes the i-th view and N is the number of input views, GeoMem3D first reconstructs multi-view observations into a geometrically consistent scene representation in a shared 3D coordinate system and then organizes the dense reconstruction into an object-centric structured 3D world memory. This two-stage design decouples 3D world perception, which estimates scene geometry and appearance, from structured memory construction, which organizes the recovered information into persistent object entities with semantic and geometric attributes and explicit spatial relations. The resulting query-independent memory provides a reusable structured representation that supports flexible, question-conditioned use of 3D information across diverse spatial reasoning tasks.

Multi-view 3D Perception and Reconstruction. Individual RGB observations provide only partial and viewpoint-dependent evidence of the underlying 3D scene. We therefore formulate multi-view reconstruction as a generic mapping $\mathcal{F}_{3D}$ that consolidates these observations into a geometri-

cally aligned scene representation:

$$\begin{aligned} \mathcal{S}_{3D} &= \mathcal{F}_{3D}(\mathcal{V}) = \{\mathcal{G}_{\text{view}}, \mathcal{P}_{\text{scene}}\}, \\ \mathcal{G}_{\text{view}} &= \{(\mathbf{K}_i, \mathbf{T}_i, \mathbf{D}_i)\}_{i=1}^{N}, \quad \mathcal{P}_{\text{scene}} = \{(\mathbf{p}_m, \mathbf{c}_m, \boldsymbol{\phi}_m)\}_{m=1}^{M}. \end{aligned} \tag{1}$$

where $\mathcal{G}_{\text{view}}$ contains the view-level geometric estimates, with $\mathbf{K}_i$, $\mathbf{T}_i$, and $\mathbf{D}_i$ denoting the camera intrinsics, camera pose, and depth map of the i-th view. These estimates are used to unproject valid depth observations into a shared reference frame and aggregate them into the scene-level point cloud $\mathcal{P}$scene. Each fused point is represented by its 3D position $\mathbf{p}_m \in \mathbb{R}^3$, RGB appearance $\mathbf{c}_m \in [0,1]^3$, and point-aligned visual feature $\boldsymbol{\phi}_m \in \mathbb{R}^C$. The resulting reconstruction provides cross-view-aligned geometry and fine-grained appearance information.

Object-centric Structured 3D World Memory. We organize the dense geometric and visual evidence in $\mathcal{S}_{3D}$ into an object-centric structured memory:

$$\mathcal{M} = \mathcal{F}_{\text{str}}(\mathcal{S}_{3D}) = (\mathcal{O}, \mathcal{R}), \tag{2}$$

where $\mathcal{O} = \{o_i\}_{i=1}^{L}$ contains L persistent object entities, and $\mathcal{R} = \{r_{ij}\}_{(i,j)\in\mathcal{E}}$ denotes the spatial relations over relevant object pairs indexed by $\mathcal{E}$. Each object is described by $o_i = (\ell_i, \mathbf{p}_i, \mathbf{s}_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}_i)$, comprising its semantic label, 3D position, spatial extent, orientation, and aggregated visual feature, respectively. These attributes consolidate the corresponding point-level observations across views into a unified object representation. Each relation is computed as $r_{ij} = \mathcal{F}_{\text{rel}}(o_i, o_j)$ and encodes geometric properties such as relative direction, distance, proximity, and containment. The resulting memory provides a compact object-relation representation of the reconstructed scene and serves as a reusable structured memory for downstream reasoning tasks.

3.3 Memory-Grounded Spatial Reasoning Framework

We treat the structured memory $\mathcal{M}$ as an explicit, query-independent spatial memory representation and formalize its use in spatial reasoning through a memory access operator:

$$\mathbf{z} = \mathcal{F}_{\text{access}}(\mathcal{M}, \mathcal{C}), \tag{3}$$

where $\mathcal{C}$ specifies the scope of memory access and $\mathbf{z}$ denotes the resulting reasoning evidence, including object attributes, spatial relations, and geometric measurements. Based on this formulation, we introduce two complementary modes: holistic reasoning over the complete memory and agentic reasoning through adaptive memory access.

Holistic World Reasoning. GeoMem-Holistic provides the MLLM with the complete structured memory in a single pass:

$$\mathbf{z}_{\text{H}} = \mathcal{F}_{\text{access}}(\mathcal{M}, \mathcal{C}_{\text{all}}), \qquad \hat{y}_{\text{H}} = f_\theta(q, \mathcal{V}, \mathbf{z}_{\text{H}}). \tag{4}$$

The MLLM selects and integrates task-relevant evidence directly from the complete structured memory.

Agentic World Reasoning. GeoMem-Agentic instead accesses the memory iteratively. At reasoning step t, the MLLM issues a memory request a_t based on the question, visual context, and current reasoning state:

$$\begin{aligned} a_t \sim \pi_\theta(\cdot \mid q, \mathcal{V}, h_t), \qquad & \mathbf{z}_t = \mathcal{F}_{\text{access}}(\mathcal{M}, a_t), \\ h_{t+1} = \mathcal{F}_{\text{reason}}(h_t, a_t, \mathbf{z}_t), \quad & \hat{y}_{\text{A}} = f_\theta(q, \mathcal{V}, h_T). \end{aligned} \tag{5}$$

where h_t summarizes the accumulated reasoning context and T denotes the step at which the MLLM determines that sufficient evidence has been acquired. Evidence selection and termination therefore constitute an adaptive world-memory usage policy rather than a fixed retrieval procedure.

3.4 Learning World-Memory Usage Policy

We introduce a tailored memory-grounded training strategy that enables the MLLM to interpret the structured 3D world encoded in memory and reason over it adaptively. The strategy first uses supervised fine-tuning on memory-grounded reasoning trajectories to teach the model how to associate spatial queries with relevant world evidence, and then applies reinforcement learning to further develop its ability to select, integrate, and reason over that evidence.

Supervised Fine-tuning with Memory-Grounded Reasoning Trajectories. We construct separate memory-grounded supervision sets for the two reasoning paradigms. The holistic dataset $\mathcal{D}_{\mathrm{H}}$ demonstrates how to identify and integrate relevant spatial knowledge from the complete structured memory. In contrast, the agentic dataset $\mathcal{D}_{\mathrm{A}}$ contains multi-step trajectories that demonstrate grounding decisions, evidence integration, and adaptive termination. We train the two models independently using their corresponding supervision:

$$\mathcal{L}_{\mathrm{SFT}}^{m} = -\mathbb{E}_{(x_m,\tau_m^*)\sim\mathcal{D}_m}\left[\log p_{\theta_m}(\tau_m^* \mid x_m)\right], \qquad m \in \{\mathrm{H},\mathrm{A}\}, \tag{6}$$

where θ_{H} and θ_{A} denote the parameters of GeoMem-Holistic and GeoMem-Agentic, respectively. This mode-specific supervision allows each model to specialize in either direct reasoning over the complete structured memory or adaptive acquisition of task-relevant world evidence.

Reinforcement Learning for Memory-Grounded Reasoning. We further optimize the memory-grounded reasoning policy using GRPO (Shao et al., 2024). For each training input $x = (q, \mathcal{V})$, the rollout policy $\pi_{\theta_{\mathrm{old}}}$ samples a group of G candidate trajectories under the same structured memory $\mathcal{M}$. These trajectories capture alternative reasoning paths, patterns of spatial evidence utilization, and final predictions, allowing the policy to be optimized according to their relative quality:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\left[\min\left(\rho_i(\theta)A_i,\ \mathrm{clip}\left(\rho_i(\theta), 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta D_{\mathrm{KL}}\left(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\right)\right], \tag{7}$$

where $\rho_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\mathrm{old}}}(o_i|q)}$ denotes the importance ratio, π_{ref} is the frozen SFT reference policy, and ϵ and β denote the clipping and KL coefficients, respectively. The group-relative advantage is computed as:

$$A_i = \frac{R(\tau_i) - \mathrm{mean}\big(\{R(\tau_j)\}_{j=1}^{G}\big)}{\mathrm{std}\big(\{R(\tau_j)\}_{j=1}^{G}\big)}, \tag{8}$$

where the mean and standard deviation are computed over the rewards of the G candidate trajectories generated for the same training instance.

Each trajectory is evaluated using a verifiable reward that captures both format compliance and task correctness:

$$\begin{aligned} R(\tau^{(g)}) &= \lambda_{\mathrm{fmt}} R_{\mathrm{fmt}}^{(g)} + \lambda_{\mathrm{cor}} R_{\mathrm{cor}}^{(g)}, \\ R_{\mathrm{cor}}^{(g)} &= \begin{cases} R_{\mathrm{acc}}^{(g)}, & q \in \mathcal{Q}_{\mathrm{cat}}, \\ R_{\mathrm{mra}}^{(g)}, & q \in \mathcal{Q}_{\mathrm{num}}, \end{cases} \end{aligned} \tag{9}$$

where R_{fmt} rewards compliance with the prescribed response format, while R_{cor} evaluates final-answer correctness. Specifically, R_{acc} measures exact correctness for categorical questions, whereas R_{mra} provides error-sensitive credit for numerical predictions. This reward encourages the model to generate well-formed reasoning trajectories while directly optimizing spatial answer quality.

4 Experiments

4.1 Experimental Settings

Baselines To comprehensively evaluate the performance of GeoMem3D, we compare it against three categories of baselines. **Closed-source MLLMs** include GPT-5-mini and Gemini-series models, including Gemini-1.5 Flash (Team et al., 2024), Gemini-1.5 Pro (Team et al., 2024), Gemini-2 Flash, and Gemini-3 Flash, representing proprietary multimodal models with general visual understanding and reasoning capabilities. **Open-source MLLMs** include representative models from the LLaVA-OV (Li et al., 2024), Qwen-VL Bai et al. (2025b;a), and InternVL Zhu et al. (2025) families. We evaluate multiple scales of Qwen3-VL (4B, 8B, and 32B) and InternVL3 variants to study the impact of model scaling on spatial reasoning. **Spatial Reasoning-enhanced Models** include Spatial-MLLM (Wu et al., 2025), SpatialLadder (Li et al., 2026a), and Think3D (Zhang et al., 2026b), which enhance spatial reasoning through specialized training, geometric modeling, or interactive 3D exploration.

Table 1: Performance comparison on VSI-Bench-Tiny. We highlight the improvements achieved by GeoMem3D through structured 3D world memory reasoning. Results are reported under the 64-frame evaluation setting, while Spatial-MLLM is evaluated with its original 16-frame setting following its best reported configuration. Think3D and GeoMem3D use the Qwen3-VL-8B backbone.

Methods	Avg.	Numerical (MRA)				Multiple-Choice (Acc.)			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Closed-source MLLMs									
GPT-5-mini	49.6	49.0	31.6	73.0	41.2	42.0	48.0	46.0	66.0
Gemini-1.5 Flash (Team et al., 2024)	45.7	50.8	33.6	56.5	45.2	48.0	39.8	32.7	59.2
Gemini-1.5 Pro (Team et al., 2024)	48.8	49.6	28.8	58.6	49.4	46.0	48.1	42.0	68.0
Gemini-2 Flash	45.4	52.4	30.6	66.7	31.8	56.0	46.3	24.5	55.1
Gemini-3 Flash	53.7	45.9	21.8	**75.0**	39.2	58.0	**62.0**	40.0	**88.0**
Open-source MLLMs									
LLaVA-OV-7B (Li et al., 2024)	29.6	30.0	27.6	39.4	22.2	29.2	30.0	35.7	23.0
LLaVA-OV-72B (Li et al., 2024)	41.4	70.0	25.2	51.9	25.2	45.8	33.7	42.9	31.2
Qwen2.5-VL-7B (Bai et al., 2025b)	32.0	17.4	28.8	59.6	25.4	25.0	33.1	35.7	31.2
Qwen2.5-VL-72B (Bai et al., 2025b)	41.4	18.2	34.3	54.6	37.0	52.1	49.3	42.9	42.6
Qwen3-VL-4B (Bai et al., 2025a)	51.3	67.7	53.9	72.7	39.3	52.1	41.7	28.5	54.1
Qwen3-VL-8B (Bai et al., 2025a)	54.9	69.8	54.6	74.1	51.6	50.0	47.8	35.7	55.7
Qwen3-VL-32B (Bai et al., 2025a)	57.4	74.3	57.9	69.7	57.2	52.1	54.6	35.7	57.4
InternVL3-8B (Zhu et al., 2025)	34.2	42.8	33.5	57.4	27.8	14.5	39.1	28.6	34.4
InternVL3-38B (Zhu et al., 2025)	43.7	75.7	36.9	60.7	31.1	47.9	38.3	21.4	37.7
InternVL3-78B (Zhu et al., 2025)	47.5	62.3	39.6	57.4	35.2	50.0	51.1	50.0	34.4
Spatial Reasoning-enhanced Models									
Spatial-MLLM (Wu et al., 2025)	45.2	67.2	31.0	56.6	43.6	32.0	52.0	36.0	44.0
SpatialLadder (Li et al., 2026a)	44.7	64.0	32.6	63.4	34.8	52.1	45.7	28.6	36.1
Think3D (Zhang et al., 2026b)	58.2	75.4	50.9	69.7	55.9	50.0	58.1	42.9	62.5
Ours									
GeoMem3D-Agentic	59.6	78.0	56.3	72.3	55.2	47.9	53.1	50.0	63.9
(Qwen3-VL-8B)	+4.7	+8.2	+1.7	-1.8	+3.6	-2.1	+5.3	+14.3	+8.2
GeoMem3D-Holistic	**63.5**	**78.2**	**58.4**	74.5	**60.2**	**58.3**	57.3	**57.1**	63.9
(Qwen3-VL-8B)	+8.6	+8.4	+3.8	+0.4	+8.6	+8.3	+9.5	+21.4	+8.2

Benchmarks We evaluate GeoMem3D on two spatial reasoning benchmarks, VSI-Bench (Yang et al., 2025) and MindCube (Wang et al., 2026). VSI-Bench is a visual-spatial intelligence benchmark based on 288 real-world indoor videos from ScanNet, ScanNet++, and ARKitScenes, containing over 5,000 QA pairs across eight spatial reasoning tasks, including object counting, metric estimation, spatial relation understanding, route planning, and approach order prediction. Following the official protocol, we report Mean Relative Accuracy (MRA) for numerical questions and Accuracy (ACC) for multiple-choice questions. MindCube evaluates spatial reasoning in controlled 3D environments through compositional queries over object attributes and spatial relationships, serving as an additional benchmark for evaluating cross-benchmark generalization.

Implementation Details GeoMem3D is trained from Qwen3-VL-8B using supervised fine-tuning followed by GRPO. We construct over 5K high-quality memory-grounded reasoning trajectories for supervised fine-tuning. Training follows the VSI-Bench setting, while MindCube-Tiny is used only for cross-benchmark evaluation. During GRPO, the rollout policy samples $G = 8$ trajectories per input. The model is trained for 1 epoch with a learning rate of 1×10^{-6} and a maximum response length of 512 tokens. The KL coefficient is set to $\beta = 0.04$. All experiments are conducted on 8 NVIDIA A100 GPUs with DeepSpeed ZeRO-3. More details can be found in the appendix.

4.2 Main Results

Comparison to State-of-the-Art Approaches. Table 1 compares GeoMem3D with existing MLLMs and spatial reasoning methods. Closed-source MLLMs achieve strong spatial reasoning performance, demonstrating the capability of large-scale multimodal models in understanding complex spatial scenes. Among open-source MLLMs, larger models generally exhibit better spatial reasoning ability, and the Qwen3-VL series achieves competitive performance across different model scales. spatial reasoning-enhanced models further improve spatial understanding through specialized training or spatially oriented modeling. Overall, these results indicate that both model scaling and spatial specialization contribute to improving spatial reasoning capability. GeoMem3D further

Table 2: Performance comparison on MindCube-Tiny. We highlight the improvements of GeoMem3D over the Qwen3-VL-8B backbone under cross-benchmark generalization.

Methods	Among ↑	Rotation ↑	Around ↑	Overall ↑
Spatial-MLLM (Wu et al., 2025)	41.8	33.5	35.6	38.8
SpatialLadder (Li et al., 2026a)	43.2	35.0	50.8	43.5
Qwen3-VL-8B (Bai et al., 2025a)	34.8	52.4	41.5	39.8
Qwen3-VL-8B-based Methods (Same Training Configuration)				
Think3D (Zhang et al., 2026b)	37.0	55.0	43.2	41.9
GeoMem3D-Agentic	38.7	**56.6**	**46.4**	43.9
(Qwen3-VL-8B)	+3.9	+4.2	+4.9	+4.1
GeoMem3D-Holistic	**41.3**	55.0	43.2	**44.4**
(Qwen3-VL-8B)	+6.5	+2.6	+1.7	+4.6

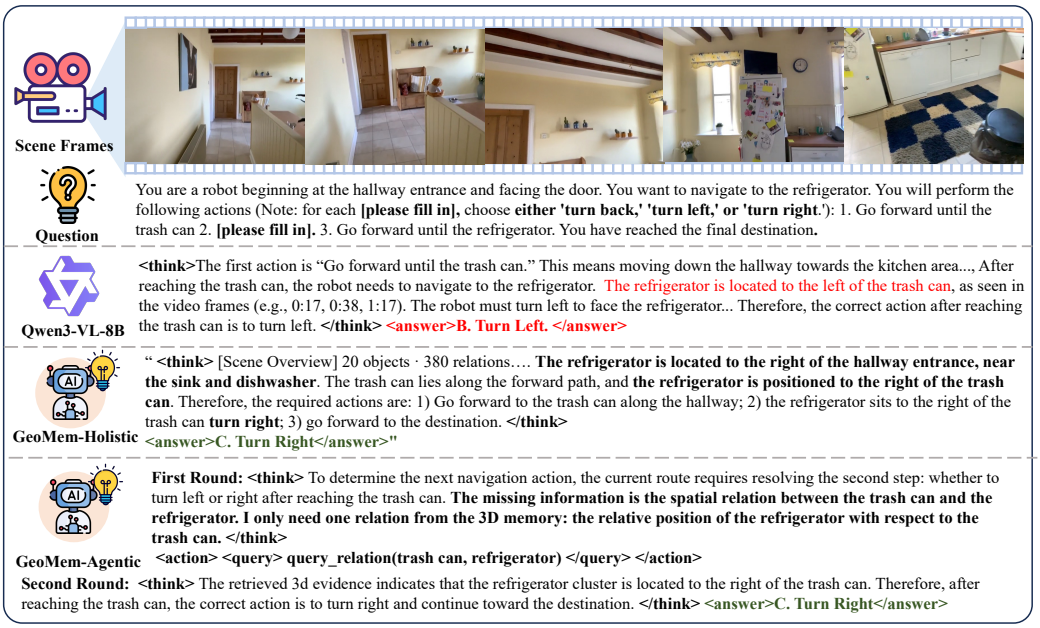


Figure 3: **Qualitative Results.** Qualitative comparison between Qwen3-VL-8B and GeoMem3D for spatial reasoning. GeoMem3D achieves more reliable spatial decisions by leveraging structured 3D world memory, while the baseline suffers from inconsistent spatial understanding across views.

advances spatial reasoning by enabling MLLMs to reason over a structured 3D world memory. Compared with existing approaches, GeoMem3D achieves the best overall performance and consistently improves spatial estimation and relational reasoning. The comparison between the two memory utilization strategies shows that both approaches improve local spatial reasoning, while GeoMem3D-Holistic provides larger gains on global scene understanding tasks such as room size estimation and route planning. This suggests that direct reasoning over a complete structured world memory is particularly beneficial when tasks require integrating information across the entire environment.

As shown in Table 2, we further evaluate the generalization capability of GeoMem3D on MindCube-Tiny (Wang et al., 2026), a cross-benchmark spatial reasoning benchmark with limited-view observations. Without additional training or adaptation on MindCube data, GeoMem3D consistently improves spatial reasoning by organizing available observations into a structured world memory. The improvements across different spatial reasoning tasks demonstrate that GeoMem3D learns transferable spatial representations beyond the original training distribution.

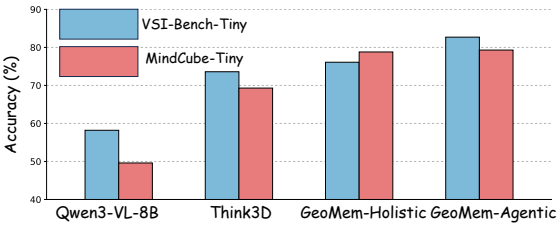


Figure 4: Qualitative analysis of model reasoning.

Table 3: Ablation study of GeoMem3D.

Setting	VSI-Bench-Tiny
Qwen3-VL-8B	54.9
+ Holistic Memory	57.4
+ Holistic Memory (SFT)	60.5
+ Holistic Memory (SFT + RL)	**63.5**
+ Agentic Memory	56.9
+ Agentic Memory (SFT)	58.7
+ Agentic Memory (SFT + RL)	59.6

Interpretability Analysis. To evaluate reasoning interpretability, we employ GPT-4o as an evaluator and assess generated explanations from three perspectives: (1) logical coherence of the reasoning process, (2) consistency between the reasoning process and the final prediction, and (3) relevance and sufficiency of the referenced spatial evidence. To decouple explanation quality from prediction, we evaluate only correctly answered instances and anonymize method identities during scoring.

As shown in Figure 4, both GeoMem3D variants generate more coherent and spatially grounded reasoning processes than Qwen3-VL-8B, demonstrating improved reasoning transparency enabled by 3D world memory. GeoMem3D-Holistic maintains consistency with the reconstructed scene by reasoning over the complete structured memory, while GeoMem3D-Agentic provides targeted reasoning traces by selectively accessing query-relevant spatial evidence. These results demonstrate that structured 3D world memory improves both spatial reasoning and reasoning transparency.

Figure 3 presents a case study illustrating how structured 3D memory resolves spatial reasoning failures caused by direct video understanding. The Qwen3-VL-8B relies on viewpoint-dependent visual observations and may produce inconsistent spatial judgments across frames. In contrast, GeoMem3D reasons over the structured memory: GeoMem-Holistic maintains globally consistent object locations and relations, while GeoMem-Agentic adaptively accesses query-relevant evidence for targeted reasoning. In this example, GeoMem3D identifies the spatial relation between the trash can and the refrigerator, retrieves the corresponding 3D evidence, and correctly infers the direction.

4.3 Ablation Studies

We evaluate the contribution of structured 3D world memory and the two-stage optimization strategy in GeoMem3D. As shown in Table 3, starting from the Qwen3-VL-8B backbone, introducing holistic or agentic memory improves spatial reasoning performance, demonstrating the benefit of structured 3D world memory. Applying supervised fine-tuning (SFT) further improves both memory utilization strategies by enabling the model to reason over the constructed 3D memory. Reinforcement learning with GRPO provides additional gains, indicating that optimizing memory-grounded reasoning behaviors further enhances spatial reasoning. Comparing the two memory utilization paradigms, holistic reasoning achieves stronger overall performance, while agentic reasoning provides an adaptive alternative through selective memory access.

5 Conclusion

In this work, we present **GeoMem3D**, a unified perception, reconstruction, memory, and reasoning framework that shifts multi-view spatial reasoning from directly interpreting viewpoint-dependent observations to reasoning over a persistent and structured 3D world memory. GeoMem3D reconstructs multi-view observations into a coherent 3D scene, organizes the recovered geometry into object-centric structured memory, and supports complementary holistic and agentic memory-grounded reasoning. A tailored two-stage learning strategy further enables MLLMs to effectively utilize this memory through supervised fine-tuning and GRPO-based reinforcement learning. Our experiments demonstrate the effectiveness and generalizability of structured 3D world memory across diverse spatial reasoning tasks and reveal complementary strengths between complete-memory reasoning and adaptive memory access. More broadly, our findings suggest that advancing spatial intelligence requires not only stronger 3D perception, but also mechanisms for organizing, preserving, and effectively utilizing structured spatial memory. Future work will extend this formulation to dynamic, uncertain, and continually evolving environments.

AI USE STATEMENT

We used generative AI tools for manuscript editing, including improving readability, grammar, and phrasing. We also used generative AI models to assist in constructing high-quality supervised fine-tuning (SFT) data. All AI-assisted data construction was carefully reviewed and verified by the authors through manual inspection and quality control procedures. No other required or recommended disclosure items apply to this work. All AI-assisted content and data were reviewed by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work aims to advance research in multimodal spatial reasoning and reliable 3D reasoning systems. GeoMem3D is developed solely for research purposes to study how MLLMs utilize structured 3D world memory for spatial understanding. All datasets are publicly available and used according to their respective licenses and intended purposes. Additional training data are generated or curated for research use with quality control and manual verification. This work does not involve human subjects, personal data collection, or private information. While GeoMem3D may support future applications in embodied intelligence, it is intended as a research framework rather than a deployment-ready system.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed framework, training procedures, and experimental settings to facilitate reproducibility. The appendix includes implementation details, model configurations, data construction procedures, and additional experimental analysis. All benchmark datasets are publicly available and evaluated following their original protocols.

REFERENCES

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

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

Jiahua Chen, Qihong Tang, Weinong Wang, and Qi Fan. Enhancing mllm spatial understanding via active 3d scene exploration for multi-perspective reasoning. *arXiv preprint arXiv:2604.06725*, 2026.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7395–7408. IEEE, 2025.

Vidisha Deshpande, Gauri Shelke, Rupesh Jaiswal, Opal Raut, and Advait Mehendale. Few-shot learning in vision-language models: A survey. In *International Conference on Computer Vision and Robotics*, pp. 115–127. Springer, 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 31054–31065, June 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20482–20494. Curran Associates, Inc., 2023. doi: 10.52202/075280-0900. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/413885e70482b95dcbeeddc1daf39177-Paper-Conference.pdf.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/680b2a8135b9c71278a09cafb605869e-Paper-Conference.pdf.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. *Advances in Neural Information Processing Systems*, 38: 67961–67988, 2026b.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

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

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 76566–76592, 2026a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Zehao Li, Hongwei Yu, Hao Jiang, Qiang Sheng, Yilong Xu, Baolong Bi, Yang Li, Zhenlong Yuan, Yujun Cai, and Zhaoqi Wang. Factguard: Agentic video misinformation detection via reinforcement learning, 2026b. URL https://arxiv.org/abs/2602.22963.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *Advances in Neural Information Processing Systems*, 38:123926–123958, 2026a.

Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, Fangqiang Ding, and Ismini Lourentzou. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation, 2026b. URL https://arxiv.org/abs/2601.07060.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68803–68832. Curran Associates, Inc., 2024. doi: 10.52202/079017-2198. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f2257d2b291b8d7e712c70b67e09412-Paper-Conference.pdf.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *Advances in Neural Information Processing Systems*, 38:140751–140774, 2026.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19446–19455. IEEE, 2025.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3960–3969, 2020. doi: 10.1109/CVPR42600.2020.00402.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Mindcube: Spatial mental modeling from limited views, 2026. URL https://arxiv.org/abs/2506.21458.

Xiaoyan Wang, Zeju Li, Yifan Xu, Jiaxing Qi, Zhifei Yang, Ruifei Ma, Xiangde Liu, and Chao Zhang. Spatial 3d-llm: Exploring spatial awareness in 3d vision-language models, 2025. URL https://arxiv.org/abs/2507.16524.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 13569–13597. Curran Associates, Inc., 2025. doi: 10.52202/085713-0457. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/142b166c8b18b43774c4f0e0c1e9948e-Paper-Conference.pdf.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098, 2025. doi: 10.1109/ICCV51701.2025.00202.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2025. URL https://arxiv.org/abs/2412.14171.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, Hongyao Tang, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 8984–9009, 2026.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 127293–127322, 2026a.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, 2025. doi: 10.1109/ICCV51701.2025.00181.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Yizhuang Peng, Zhenfei Yin, Lijun Wang, and Huchuan Lu. Think3d: Thinking with space for spatial reasoning, 2026b. URL https://arxiv.org/abs/2601.13029.

Linqing Zhong, Chen Gao, Zihan Ding, Yue Liao, Huimin Ma, Shifeng Zhang, Xu Zhou, and Si Liu. Topv-nav: Unlocking the top-view spatial reasoning potential of mllm for zero-shot object navigation. *International Journal of Computer Vision*, 134(9):409, 2026.

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

A Appendix

In this appendix, we provide more experimental details and method details. Detailed contents are as follows:

A.1 More Experimental Details

Datasets. VSI-Bench (Yang et al., 2025) is a visual-spatial intelligence benchmark designed to evaluate the ability of multimodal large language models to perform spatial reasoning in realistic 3D environments. Unlike conventional image-based visual question answering benchmarks, VSI-Bench focuses on spatial understanding under partial and viewpoint-dependent observations, requiring models to infer geometric relationships beyond individual frames.

The benchmark is constructed from 288 real-world indoor videos collected from ScanNet, ScanNet++, and ARKitScenes, covering diverse indoor environments with complex object arrangements and camera viewpoints. It contains over 5,000 question-answer pairs spanning eight spatial reasoning tasks: object counting, absolute distance estimation, object size estimation, room size estimation, relative distance reasoning, relative direction reasoning, route planning, and approach order prediction.

These tasks evaluate different aspects of spatial intelligence, including metric estimation, relational reasoning, and navigation-oriented decision making. Following the official evaluation protocol, numerical questions are evaluated using Mean Relative Accuracy (MRA), while multiple-choice questions are evaluated using Accuracy (ACC). We follow the original benchmark settings for all VSI-Bench experiments.

MindCube (Wang et al., 2026) is a 3D spatial reasoning benchmark designed to evaluate whether vision-language models can build and utilize spatial mental models from limited visual observations. Compared with real-world video benchmarks, MindCube focuses on controlled spatial reasoning scenarios with well-defined spatial configurations, enabling evaluation of fine-grained reasoning over object attributes and spatial relationships.

The benchmark contains queries that require models to reason about object properties, relative positions, and spatial configurations. Such controlled settings reduce visual ambiguity and focus evaluation on the ability to infer structured spatial relationships.

Training Details GeoMem3D is initialized from the Qwen3-VL-8B-Instruct. To establish memory-grounded spatial reasoning ability, we construct over 5K high-quality supervised fine-tuning trajectories. We employ Qwen3-VL-32B as the teacher model to generate candidate reasoning trajectories based on structured 3D world memory and spatial queries. For each sample, the corresponding ground-truth answer is provided as an auxiliary guidance signal to improve generation reliability and facilitate quality verification, rather than being used to derive or constrain the intermediate reasoning process. Candidate trajectories are further filtered through automatic validation and manual inspection, removing malformed outputs, inconsistent reasoning traces, and incorrect predictions. The resulting high-quality trajectories are used for supervised fine-tuning.

Supervised fine-tuning is performed on the constructed memory-grounded reasoning trajectories before reinforcement learning. Training is conducted with BF16 precision and DeepSpeed ZeRO-2 optimization. The model is trained for 3 epochs with a learning rate of 1×10^{-5} (cosine schedule, no warm-up), batch size of 1 per device, and gradient accumulation of 2 across 8 GPUs. Gradient checkpointing is applied throughout training, and FlashAttention is enabled where supported.

Following supervised fine-tuning, we further optimize GeoMem3D with GRPO (Shao et al., 2024). For each input, the rollout policy samples $G = 8$ candidate trajectories, enabling optimization over different memory-grounded reasoning behaviors. The reward function follows Equation 9, combining response format compliance and spatial answer correctness. The model is trained for 1 epoch with a learning rate of 1×10^{-6}. The KL coefficient is set to $\beta = 0.04$, and the maximum response length is set to 512 tokens. All reinforcement learning experiments are conducted on 8 NVIDIA A100 GPUs with DeepSpeed ZeRO-3. Additional implementation and training details are provided in the supplementary material.

A.2 Additional Method Details

Multi-view 3D Reconstruction. To construct the structured 3D world memory, GeoMem3D first reconstructs the observed environment from multi-view RGB inputs. We adopt VGGT as the 3D perception front-end to estimate view-level geometric information, including camera parameters, depth, and cross-view aligned scene representations. The resulting geometric and appearance information is then organized into object-centric structured memory through the proposed memory construction module. Importantly, GeoMem3D decouples 3D perception from structured memory construction through a unified interface, making the framework independent of any specific reconstruction model. Therefore, alternative 3D reconstruction or perception models, such as Pi3 or future 3D foundation models, can be seamlessly integrated as interchangeable front-ends while keeping the downstream memory representation and reasoning process unchanged. While the memory provides an object-centric abstraction for efficient reasoning, it preserves geometric grounding through explicit attributes, spatial relations, and associated visual evidence.

Memory Representation and Query Design. Our structured 3D world memory organizes reconstructed scene information into an object-centric representation containing object entities, geometric attributes, and spatial relations. Given the reconstructed multi-view scene, we first associate observations across views and aggregate them into persistent object entities. Each entity stores semantic and geometric attributes, including its centroid, 3D bounding box, depth statistics, dominant appearance, visibility information, and reconstruction confidence, together with the corresponding visual evidence for later verification. Spatial relations are represented as typed directed connections between object entities, capturing geometric properties such as relative distance, directional relations, proximity, and containment. This representation transforms dense reconstruction into a compact and reusable memory structure, where object-level information and relational knowledge can be accessed independently of individual observations.

To enable flexible memory utilization, we design a unified query interface that retrieves structured spatial evidence rather than raw geometric data. Unlike conventional retrieval mechanisms over unstructured information, these queries operate on explicit 3D entities and relations, providing geometrically grounded evidence for reasoning. Object-level queries retrieve entity-specific semantic and geometric attributes through object identifiers, textual descriptions, or visual references. Relation-level queries access spatial relationships between object pairs, including relative direction, distance, proximity, and containment. Region-level queries retrieve objects and spatial information within specified areas of the scene, supporting reasoning over local configurations. Measurement-

level queries provide precise geometric comparisons, such as distances and spatial extents between selected entities. The retrieved evidence is converted into structured representations and natural language descriptions, allowing MLLMs to directly incorporate memory information during reasoning.

These query mechanisms enable GeoMem-Agentic to adaptively access task-relevant spatial evidence during reasoning, selecting appropriate memory granularity according to the query and intermediate reasoning state. In contrast, GeoMem-Holistic reasons over a complete serialization of the structured 3D world memory in a single pass. Both paradigms operate on the same underlying memory representation and query interface, allowing us to study different strategies for memory utilization while keeping the stored spatial knowledge unchanged.