THINKING WITH LATENT GEOMETRY TOKENS FOR MULTI-IMAGE SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-image spatial reasoning remains challenging for multimodal large language models (MLLMs). A key bottleneck lies in integrating partial observations into coherent internal spatial representations, a capability known as spatial mental modeling. Existing approaches often rely on dense spatial annotations or external tools, limiting scalability or increasing inference overhead. To reduce these dependencies, recent work explores modeling spatial representations in latent space. However, effectively encoding geometric information in latent space and incorporating it during reasoning is underexplored. We propose **Thinking with Latent Geometry Tokens**, a two-stage framework that grounds latent states in compact geometric representations and encourages flexible latent-mode entry. Specifically, we use camera and scene tokens from a pretrained 3D foundation model to supervise latent states, providing viewpoint-conditioned and cross-view geometry information while jointly training on text generation. We then augment Group Relative Policy Optimization (GRPO) with latent-mode entry point exploration, using successful exploratory trajectories to guide policy updates. Experiments on three multi-image benchmarks demonstrate our consistent gains over spatial reasoning baselines, with further improvements in single-image settings.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved remarkable progress across fundamental visual tasks (Bucciarelli et al., 2024; Li et al., 2024a; An et al., 2025), yet reliable spatial reasoning remains elusive. Despite advances in single-image spatial tasks (Chen et al., 2024a), current MLLMs still struggle with multi-image spatial reasoning (Yin et al., 2025; Yang et al., 2026a), where evidence is distributed across partial views. Recent studies (Yin et al., 2025; Wang et al., 2026b) suggest that a key bottleneck lies in their limited capacity for spatial mental modeling (Johnson-Laird, 1980), the ability to consolidate local observations into a coherent, viewpoint-consistent internal representation of the environment. This limitation hinders the reliable deployment of MLLMs in spatially demanding applications such as robotic navigation (Qiao et al., 2026), manipulation (Li et al., 2024b), and augmented-reality interaction (Chandrasegaran et al., 2024).

Recent efforts enhance spatial mental modeling through structured spatial supervision or external spatial tools. Supervision-based methods strengthen reasoning with cognitive maps (Deng et al., 2026; Yin et al., 2025; Choi et al., 2026), cross-view correspondences (Oi et al., 2026; Wang et al., 2026d; 2025b), and viewpoint-transition action sequences (Oi et al., 2026). Tool-based methods (Zhang et al., 2026d; Luo et al., 2026) augment reasoning by reconstructing 3D scenes with 3D foundation models such as VGGT (Wang et al., 2025a) and exploring them through novel-view rendering. However, constructing structured supervision often requires scene annotations or task-specific metadata, which can constrain scalability, while external tool execution introduces additional inference overhead. These limitations motivate a central research question: *how can we elicit MLLMs' intrinsic capacity for spatial mental modeling, without relying on spatial annotations for training or external tools at inference?*

Latent visual reasoning (Li et al., 2026a; Xu et al., 2026; Yang et al., 2026b) offers a promising starting point by interleaving text generation with continuous latent states supervised by visual signals. This suggests an opportunity to construct and reason over internal spatial representations in latent space, yet how to realize such spatial mental modeling remains an open challenge. Recent

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: Overview. (a) A multi-image spatial reasoning example comparing our method with a prior spatial mental modeling method and the base model. Our method relates observations across camera viewpoints and correctly identifies the target object. (b) Our method achieves consistent gains on multi-image benchmarks, with improvements extending to single-image benchmarks despite using only multi-image data for training.

work (Chen et al., 2026) supervises latent states with dense geometric features from VGGT and enters latent mode at a fixed point before text generation. We empirically find this formulation suboptimal for two reasons: **(i)** latent states remain highly similar across instances, suggesting that current supervision may not sufficiently encourage latent states to encode input-specific spatial information; and **(ii)** effective latent-mode entry points vary across instances, indicating that a fixed entry point may limit reasoning performance.

To bridge this gap, we introduce **Thinking with Latent Geometry Tokens**, a two-stage framework that equips MLLMs with spatial mental modeling in latent space. The core idea is to align latent states with compact geometry tokens and enable flexible latent-mode entry as reasoning unfolds. In the first stage, supervised finetuning (SFT) aligns latent states with compact camera tokens and scene tokens extracted from a pretrained 3D foundation model, while jointly optimizing next-token prediction for text generation. In the second stage, we employ Group Relative Policy Optimization (GRPO) to effectively leverage these geometry-grounded latent representations during reasoning. Instead of relying solely on standard rollouts, we explore alternative latent-mode entry points when none of the initial rollouts yields a correct answer. Successful exploratory trajectories are incorporated into policy updates, providing additional learning signals.

Our method consistently outperforms baselines on multi-image benchmarks while demonstrating strong generalization to single-image settings, as shown in Figure 1. We further validate its applicability across model families and scales. Our main contributions are as follows:

- We introduce **Thinking with Latent Geometry Tokens**, a simple yet effective framework that equips MLLMs with spatial mental modeling in latent space by grounding latent states in compact geometry tokens and enabling flexible latent-mode entry.
- We propose a two-stage training strategy that aligns latent states with compact camera tokens and scene tokens through supervised finetuning and explores alternative latent-mode entry points during reinforcement learning.
- Extensive experiments demonstrate consistent gains of our method across multi-image spatial reasoning benchmarks and show superior performance on single-image tasks. We further demonstrate the broad applicability across various model families and scales.

2 Method

2.1 Preliminaries

Problem Formulation. Given a set of images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ depicting a physical scene and a question q, multi-image spatial reasoning aims to generate a textual response $y = (y_1, \ldots, y_{T_y})$ by

integrating evidence across views. Following latent visual reasoning (Li et al., 2026a), we consider a generation process that interleaves discrete text tokens with continuous latent states. The complete reasoning trajectory o is represented as

$$o = y_{1:k} \oplus \mathbf{z}_{1:K} \oplus y_{k+1:T_y}, \qquad 0 \le k < T_y, \tag{1}$$

where $\oplus$ denotes sequence concatenation, T_y denotes the number of text tokens, k denotes the latent-mode entry position, and $\mathbf{z}_{1:K}$ comprises K continuous latent embeddings.

Visual Geometry Grounded Transformer. VGGT (Wang et al., 2025a) is a feed-forward model that directly predicts various 3D attributes of a scene from 2D images. To integrate information within and across views, its transformer backbone interleaves frame-wise self-attention and global self-attention. Building on that, VGGT-Ω (Wang et al., 2026a) replaces a subset of global attention layers with attention among register (scene) tokens. These registers exchange information across views and redistribute it to image tokens through subsequent frame-wise attention. For each view, the model maintains a camera token trained for camera-parameter prediction and register tokens for aggregating scene context across views. Their utility beyond reconstruction is demonstrated by successful text alignment with scene descriptions through contrastive finetuning and improved robotic manipulation performance when used as additional inputs to a vision-language-action model (Wang et al., 2026a). In our work, we use these compact representations to ground latent states in viewpoint-conditioned and cross-view geometric information.

2.2 Motivation

To conduct analysis, we use 3DThinker (Chen et al., 2026) as a representative implementation which supervises latent states with dense geometry features and enters latent mode at a fixed point before text generation. First, we compute pairwise cosine similarities between latent states at the same latent index across distinct instances randomly sampled from MindCube (Yin et al., 2025) and MMSI-Bench (Yang et al., 2026a). As illustrated in Figure 2(a), latent states exhibit high similarity for most instance pairs, suggesting limited differentiation across diverse inputs. This observation motivates investigating whether dense geometric supervision sufficiently encourages latent states to encode input-specific spatial information. We hypothesize that dense geometric supervision may be poorly matched to the granularity of latent states, and therefore explore compact geometric representations as alternative supervision targets.

Second, we investigate whether changing the latent-mode entry points can help the model solve previously unsuccessful cases. We sample 100 hard cases that the model fails under its default setting and revisit them with alternative entry points. All runs use greedy decoding. As shown in Figure 2(b), changing the entry point corrects several previously failed cases, and the cases recovered at different positions only partially overlap. These observations suggest that a fixed latent-mode entry point may limit the model's ability to solve challenging questions.

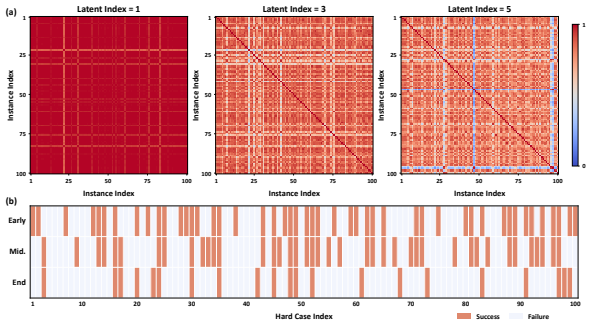


Figure 2: (a) Pairwise cosine similarities between latent states at the same latent index. (b) Answer correctness under alternative latent-mode entry points. Early and Mid. place entry at approximately 25% and 50% of the average output length in tokens, respectively. End places entry immediately before the answer tag.

2.3 Thinking with Latent Geometry Tokens

The two-stage framework of Thinking with Latent Geometry Tokens is in Figure 3. In the first stage, supervised finetuning aligns latent states with compact camera and scene tokens from a pretrained VGGT-Ω model. These tokens provide complementary viewpoint-conditioned and cross-view geometric information. By jointly optimizing geometric alignment and next-token prediction, we train

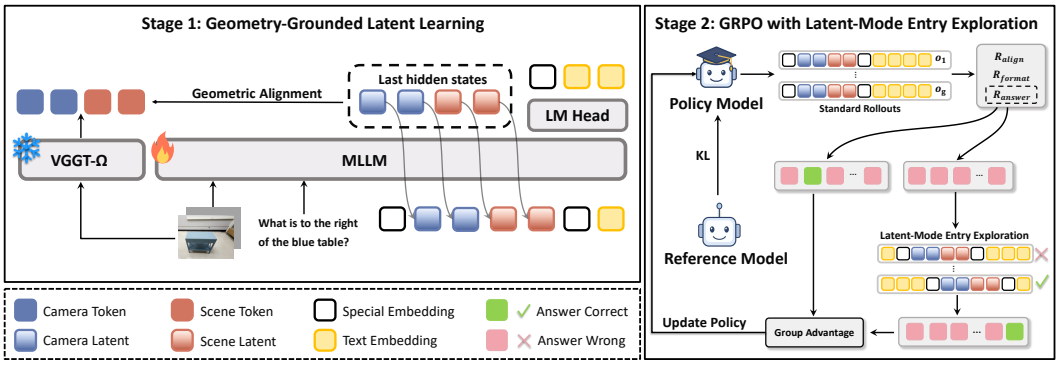


Figure 3: Overall framework of Thinking with Latent Geometry Tokens. Given multiple images and a question, Stage 1 grounds latent states in compact camera and scene tokens extracted from a frozen VGGT-Ω model, jointly optimizing geometric alignment and next-token prediction. Stage 2 refines the policy using GRPO with latent-mode entry exploration. When none of the initial rollouts yields a correct answer, alternative entry points along the reasoning trajectory are explored. The successful exploratory trajectories are combined with the initial rollouts for policy optimization.

the model to encode geometric information in latent states and use it for subsequent textual reasoning. In the second stage, we employ GRPO to further improve reasoning with the learned latent representations. When standard rollouts yield no correct answers, we explore alternative latent-mode entry points along the reasoning trajectory. Successful exploratory trajectories are incorporated into policy updates, providing additional learning signals.

2.3.1 Stage 1: Geometry-Grounded Latent Learning

The first stage aims to learn latent representations that encode complementary geometric information to support subsequent reasoning. We follow LVR (Li et al., 2026a) to adopt a teacher-forcing paradigm in which the MLLM's final-layer hidden states are supervised to reconstruct geometric targets. Each training example consists of multiple input images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, a question q, and a reference textual response y. We first feed the images $\mathcal{I}$ into a pretrained VGGT-Ω model and extract camera and register tokens from its final aggregation layer. For each view I_i, we retain its camera token $\mathbf{g}_i^{\text{cam}}$ and average its register tokens to obtain a scene token $\mathbf{g}_i^{\text{scene}}$. Each target is mapped into the language model's input embedding space by a shared learnable projector P, with a learnable type embedding $\mathbf{e}_\tau$ distinguishing the two token types:

$$\mathbf{x}_i^\tau = P(\mathbf{g}_i^\tau) + \mathbf{e}_\tau, \qquad \tau \in \{\text{cam}, \text{scene}\}, \tag{2}$$

where $\mathbf{x}_i^\tau$ denotes the corresponding input embedding. The resulting embeddings are enclosed by the special tokens <geo_start> and <geo_end> to form a geometry-grounded latent segment, followed by the reference textual response. We train the model with two joint learning objectives that explicitly couple geometry-grounded latent reasoning with text generation.

Geometric Alignment Loss. During latent mode, we supervise the model's final-layer hidden states to encode camera and scene information. Let $\mathbf{h}_i^\tau$ denote the hidden state immediately preceding the teacher-forced input $\mathbf{x}_i^\tau$. We use two separate two-layer MLP heads H_τ to map these states into the geometric feature space and supervise the predictions using a mean squared error (MSE) objective:

$$\mathcal{L}_{\text{geo}} = \frac{1}{N d_g} \sum_{i=1}^{N} \sum_{\tau} \| H_\tau(\mathbf{h}_i^\tau) - \mathbf{g}_i^\tau \|_2^2 , \tag{3}$$

where $\mathbf{g}_i^\tau$ denotes the corresponding geometric target, and d_g is the geometric feature dimension.

Next-Token Prediction Loss. We supervise discrete output tokens $y = \{y_t\}_{t=1}^{T_y}$, including textual response and the special tokens <geo_start> and <geo_end>, using standard cross-entropy loss:

$$\mathcal{L}_{\text{NTP}} = -\frac{1}{T_y} \sum_{t=1}^{T_y} \log p_\theta(y_t \mid \mathcal{I}, q, o_{<t}) , \tag{4}$$

where T_y denotes the number of the discrete output tokens and $o_{<t}$ denotes the preceding sequence, including the latent embeddings and earlier discrete tokens.

Overall Training Objective. The overall training objective of the first stage combines geometric alignment loss with next-token prediction loss:

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{NTP}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}, \tag{5}$$

where λ_{geo} balances geometric alignment and generation signals.

Model Inference. During latent reasoning, the model autoregressively predicts geometric features and feeds them back through the shared input projector with type embeddings. After $2N$ latent steps, the model resumes text generation. VGGT-Ω is not required at inference.

2.3.2 Stage 2: GRPO with Latent-Mode Entry Exploration

The SFT stage establishes the model's ability to encode geometric information in latent states and use it for textual reasoning. Building on this foundation, we employ Group Relative Policy Optimization (GRPO) with latent-mode entry exploration to enable learning from trajectories with diverse entry points. Our core idea is to preserve effective reasoning behavior on questions the model can already solve while providing additional learning signals for difficult ones. To this end, we explore alternative entry points when standard rollouts yield no correct answers and incorporate successful exploratory trajectories into policy updates.

Reward Design. We evaluate each trajectory using answer correctness, format validity, and latent geometric alignment. We compute the answer reward R_{answer} using exact option matching for multiple-choice questions and mean relative accuracy (MRA) for numerical questions. The binary format reward R_{format} encourages well-formed outputs, assigning 1 to responses with properly paired `<geo_start>` and `<geo_end>` tokens and well-formed `<think>...</think>` and `<answer>...</answer>` blocks, and 0 otherwise. The latent alignment reward measures the cosine similarity between predicted geometric features and their corresponding VGGT-Ω targets:

$$R_{\text{align}}(o) = \frac{1}{2N}\sum_{i=1}^{N}\sum_{\tau}\cos\left(\hat{\mathbf{g}}_i^{\tau}, \mathbf{g}_i^{\tau}\right), \tag{6}$$

where $\hat{\mathbf{g}}_i^{\tau} = H_{\tau}(\mathbf{h}_i^{\tau})$ is the predicted camera or scene feature. The combined reward is

$$R(o) = R_{\text{answer}}(o) + R_{\text{format}}(o) + \lambda_{\text{align}}R_{\text{align}}(o). \tag{7}$$

These complementary signals encourage correct, well-formed responses while maintaining geometric alignment of the latent representations.

Latent-Mode Entry Exploration. For each image–question pair $(\mathcal{I}, q)$, we sample initial rollouts without exploration bias. A rollout is considered successful if its predicted option matches the reference for multiple-choice questions, or its MRA score is at least 0.5 for numerical questions. If none of the initial rollouts satisfies this criterion, we sample additional trajectories by adding a position-dependent bias to the logit of `<geo_start>` before temperature scaling. We partition a predefined token range into equal intervals and uniformly sample one bias-transition center from each interval, assigning one to each exploratory trajectory. Before the first latent-mode entry, each trajectory uses a shared capped linear bias schedule shifted to its assigned center, initially discouraging immediate entry and progressively encouraging later entry. Exploratory trajectories satisfying the same success criterion are retained alongside the initial rollouts for policy optimization.

Policy Optimization. Let $\mathcal{G} = \{o_j\}_{j=1}^{G}$ contain the initial rollouts and retained successful exploratory trajectories. We adapt GRPO to this group by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{j=1}^{G}\frac{1}{|\mathcal{T}_j|}\sum_{t\in\mathcal{T}_j}\Big[\min\left(\rho_{j,t}\hat{A}_j,\ \text{clip}(\rho_{j,t}, 1-\epsilon, 1+\epsilon)\hat{A}_j\right) - \beta D_{j,t}\Big], \tag{8}$$

Table 1: Quantitative results on multi-image spatial reasoning benchmarks, including MindCube-Tiny, MMSI-Bench, and SPAR-Bench-MV. **Bold** and underlined values indicate the best and second-best results among Qwen2.5-VL based spatial models, respectively. Proprietary models and open-weight models are included for reference and excluded from this ranking. Abbreviations: Rot. = Rotation; Amo. = Among; Aro. = Around; Pos. = Position; Att. = Attribute; Mot. = Motion; MSR = Multi-step Reasoning; Mid. = Middle.

Model	MindCube-Tiny				MMSI-Bench					SPAR-Bench-MV				Overall
	Rot.	Amo.	Aro.	Avg.	Pos.	Att.	Mot.	MSR	Avg.	Low	Mid.	High	Avg.	
Proprietary Models														
GPT-4.1	49.5	47.2	59.6	50.6	29.3	33.8	38.7	31.3	31.7	38.2	37.0	43.1	39.9	40.7
GPT-5	94.5	38.2	68.4	56.1	45.0	45.9	34.6	36.4	41.9	36.3	38.9	60.5	47.0	48.3
Gemini-3-Pro	93.0	60.7	77.2	70.8	46.5	51.6	44.7	37.9	45.2	34.5	36.7	59.7	45.6	53.9
Open-Weight Models														
LLaVA-OneVision-4B	33.5	39.2	50.0	40.7	30.8	26.9	26.0	26.8	28.8	23.0	28.0	40.1	31.4	33.6
InternVL-2.5-4B	33.5	45.2	57.6	45.9	28.7	26.2	20.7	28.3	27.1	30.3	30.1	38.5	33.6	35.6
InternVL-2.5-8B	37.0	20.3	25.6	24.8	31.0	30.0	20.7	23.2	27.8	31.8	30.8	48.2	38.3	30.3
Qwen2.5-VL-32B	41.5	38.2	43.6	40.1	25.9	26.9	28.7	27.3	26.7	25.6	29.0	45.4	34.7	33.9
Qwen2.5-VL-72B	42.0	41.2	44.8	42.2	33.3	38.5	24.7	28.8	31.8	26.9	32.7	43.6	35.4	36.5
Qwen2.5-VL-7B Based Spatial Models														
Qwen2.5-VL-7B	33.5	32.7	38.0	34.1	28.4	21.5	29.3	25.8	27.1	16.9	29.4	40.7	30.2	30.5
SpaceR-7B	31.5	28.7	31.6	29.9	30.3	25.4	22.0	21.7	26.7	31.6	32.3	49.2	39.1	31.9
Video-R1	30.0	38.5	44.0	38.2	30.5	33.8	18.7	24.7	28.0	29.4	30.0	46.2	36.5	34.2
Qwen2.5-VL-3B Based Spatial Models														
Qwen2.5-VL-3B	34.5	35.3	46.4	37.8	27.8	19.2	23.3	25.8	25.6	13.4	26.2	32.6	24.9	29.4
Spatial-MLLM-4B	29.0	34.5	52.0	37.6	28.0	21.5	14.7	22.7	24.1	22.3	32.4	35.2	30.4	30.7
3DThinker-3B	44.0	64.8	72.4	62.7	20.9	12.3	21.3	16.7	19.0	25.8	17.6	32.6	26.4	36.0
SpatialLadder-3B	31.5	47.8	56.4	46.8	24.1	23.1	26.7	20.2	23.6	26.2	33.5	44.5	35.8	35.4
HATCH-3B	31.0	50.2	65.6	50.2	30.5	20.8	23.3	24.7	27.0	41.3	**47.4**	67.1	53.7	43.6
Ours-3B	**69.5**	**77.8**	**84.4**	**77.8**	**37.4**	**34.6**	**33.3**	**29.3**	**34.8**	**43.1**	44.9	**68.6**	**54.1**	**55.6**

where $\hat{A}_j = \frac{r_j - \text{mean}(r_1,\ldots,r_G)}{\text{std}(r_1,\ldots,r_G)+\delta}$ is the group-relative advantage, r_j denotes the combined reward for trajectory o_j, and $\mathcal{T}j$ indexes sampled discrete output positions. The probability ratio $\rho j, t$ compares the current and rollout policies under the same sampling bias, and Dj, t is the KL divergence between the current policy and the fixed reference model without entry bias. Here, ϵ, β, and $\delta > 0$ specify the clipping threshold, KL weight, and numerical stabilizer, respectively.

3 Experiments

3.1 Experimental Setup

Evaluation Benchmarks. We evaluate our method on four spatial reasoning benchmarks across multi-image and single-image settings. For multi-image evaluation, we use MindCube-Tiny (Yin et al., 2025), MMSI-Bench (Yang et al., 2026a), and SPAR-Bench-MV (Zhang et al., 2026a). MindCube-Tiny assesses spatial mental modeling from limited views, including perspective taking and reasoning under partial observability. MMSI-Bench evaluates multi-image spatial reasoning that requires integrating evidence across various viewpoints or time. SPAR-Bench-MV is the multi-image subset of SPAR-Bench (Zhang et al., 2026a) that evaluates spatial perception and reasoning across multiple views. To assess generalization to single-image settings, we further evaluate on SPAR-Bench-SI, which is the single-image subset of SPAR-Bench, and CV-Bench (Tong et al., 2024) for 2D and 3D visual understanding. We follow the official evaluation protocols, reporting accuracy for multiple-choice questions and mean relative accuracy (MRA) for numerical questions in SPAR-Bench. Additional details of benchmarks are provided in the Appendix section A.

Baselines. We compare our method with recent spatial reasoning approaches, including SpaceR (Ouyang et al., 2025), Video-R1 (Feng et al., 2026), Spatial-MLLM (Wu et al., 2026), 3DThinker (Chen et al., 2026), SpatialLadder (Li et al., 2026d), and HATCH (Oi et al., 2026). HATCH structures spatial reasoning through explicit intermediate supervision of cross-view corre-



Figure 4: Qualitative comparison with Qwen2.5-VL-3B. The attention maps and results are shown for both models, with attention differences in the rightmost column.

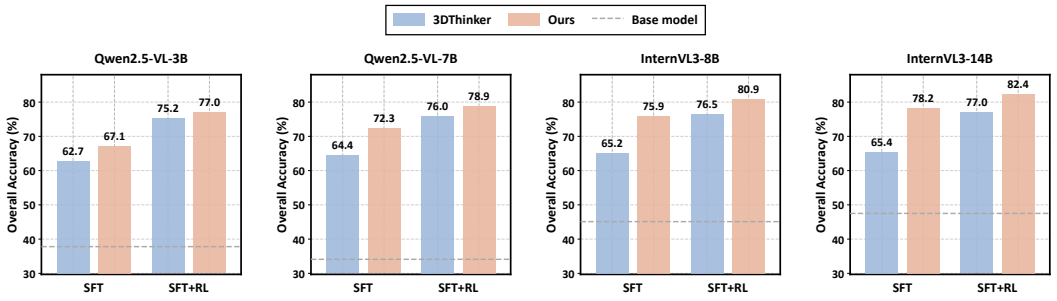


Figure 5: Comparison with 3DThinker on MindCube-Tiny across four backbone models trained on the 3DThinker-10K dataset.

spondences and viewpoint-transition actions. We adopt it as our primary baseline. 3DThinker aligns latent states with dense geometric features for spatial reasoning. We further compare with it across model families and scales. For a broader comparison, we also consider proprietary models including GPT-4.1 (Hurst et al., 2024), GPT-5 (Singh et al., 2025), and Gemini-3-Pro (Google, 2025), and open-weight models from the InternVL-2.5 (Chen et al., 2024b), LLaVA-OneVision-4B (An et al., 2025), and Qwen2.5-VL (Bai et al., 2025) families.

Implementation Details. We use Qwen2.5-VL-3B as the default backbone. In both stages, we freeze the visual encoder, pretrained multimodal projector, and VGGT-Ω. During SFT, we update all remaining model parameters. During RL, we additionally freeze the geometric prediction heads and latent input projector. We sample 15,000 multi-image instances from MindCube-Train (Yin et al., 2025) and SPAR-7M (Zhang et al., 2026a), allocating 10,000 to SFT and 5,000 to RL. Reasoning chains are curated with Qwen3.5-122B-A10B (Team, 2026) using prompts from 3DThinker (Chen et al., 2026). SFT runs for 5 epochs with a batch size of 16, a learning rate of 1×10^{-5}, and $\lambda_{\text{geo}} = 0.1$. RL runs for 2 epochs with a learning rate of 5×10^{-6}, $\beta = 0.04$, $\epsilon = 0.2$, and $\lambda_{\text{align}} = 0.1$. Standard and exploratory rollout groups each contain 8 samples, generated at temperature 0.9 with a maximum completion length of 512 tokens. We use entry-logit bias ranging from -16 to 16. Transition centers are sampled uniformly, one from each of eight equal intervals in $[32, 256)$. All experiments are conducted on 4 NVIDIA H200 GPUs. Additional details are provided in the Appendix section B.

3.2 Main Results in Multi-Image Spatial Reasoning

Quantitative Results. Table 1 summarizes the results on three multi-image spatial reasoning benchmarks. Our method achieves the highest overall score of 55.6 among the compared models, exceeding the base model Qwen2.5-VL-3B by 26.2 points and HATCH by 12.0 points. The gains over HATCH are particularly pronounced on MindCube-Tiny (+27.6 points) and MMSI-Bench (+7.8 points), while improving on SPAR-Bench-MV from 53.7 to 54.1. Despite using a 3B backbone, our

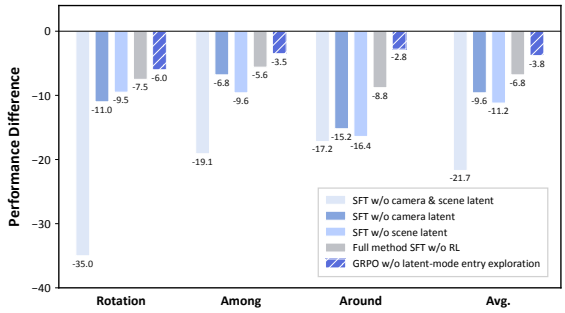


Figure 6: Ablation study on MindCube-Tiny. Bars show performance differences in percentage points relative to the full method.

Model	SPAR-Bench-SI Low	High	Avg.	CV-Bench 2D	3D	Avg.
Qwen2.5-VL-7B Based Spatial Models						
Qwen2.5-VL-7B	18.7	45.2	31.4	74.8	83.3	79.1
SpaceR-7B	33.6	46.3	39.7	71.9	81.1	76.4
Video-R1	33.8	42.7	38.1	71.7	74.9	73.3
Qwen2.5-VL-3B Based Spatial Models						
Qwen2.5-VL-3B	16.0	32.5	23.9	69.3	72.2	70.7
Spatial-MLLM-4B	24.3	37.6	30.7	65.7	69.5	67.6
3DThinker-3B	25.9	34.3	30.0	41.5	53.8	47.7
SpatialLadder-3B	25.7	39.8	32.5	72.2	74.6	73.4
HATCH-3B	41.1	49.7	45.2	70.5	**78.4**	74.5
Ours-3B	**45.5**	**50.1**	**47.7**	**74.5**	77.3	**75.8**

Table 2: Single-image spatial reasoning results. **Bold**/underlined values mark the best/second-best 3B-based results.

method also outperforms the compared 7B-based spatial reasoning models and larger open-weight models, including Qwen2.5-VL-72B, in average performance on all three benchmarks. Compared with proprietary models, it surpasses Gemini-3-Pro on MindCube-Tiny and SPAR-Bench-MV, although a gap remains on MMSI-Bench (34.8 vs. 45.2). These results demonstrate the effectiveness of our method across different multi-image spatial reasoning tasks.

Qualitative Analysis. Figure 4 compares our method with the base model on two spatial reasoning examples. In the first case, the target object is outside the reference image, requiring the model to relate observations from other views to the reference camera's orientation. Our method correctly identifies the display shelves and shows greater attention to their region in another view. Camera-token supervision may support this cross-view interpretation. In the second case, the seating bench is visible in the reference image but occupies a small peripheral region. The base model incorrectly selects the wall with an air conditioning unit, whereas our method answers correctly and assigns greater attention to the bench. Scene-token supervision may help the model incorporate these easily overlooked cues through aggregated scene context. These observations are consistent with the intended roles of camera and scene tokens.

Performance across Backbones. Figure 5 compares our method with 3DThinker on MindCube-Tiny across four backbones: Qwen2.5-VL-3B/7B and InternVL3-8B/14B. Both methods are trained on 3DThinker-10K (Chen et al., 2026). Our method consistently outperforms 3DThinker across all four backbones at both the SFT and SFT+RL stages. After SFT, the improvements range from 4.4 to 12.8 points, showing that the benefits emerge before reinforcement learning. Following RL, our method maintains gains of 1.8 to 5.4 points, achieving its highest accuracy of 82.4% with InternVL3-14B, compared with 77.0% for 3DThinker. These consistent improvements support the applicability of our framework across broader model families and parameter scales.

3.3 Ablation Studies

Component Analysis. Figure 6 summarizes component ablations on MindCube-Tiny under the training setup of Table 1. Relative to full SFT (71.0%), removing camera or scene latents reduces overall accuracy by 2.8 and 4.4 percentage points, respectively, supporting their joint use. Removing the entire latent segment (plain-text SFT) leads to a larger drop of 14.9 points. Omitting RL reduces accuracy by 6.8 points relative to the full method (77.8%). For the latent-mode entry exploration ablation, we match the additional rollout budget. When all initial rollouts fail, the ablated model samples additional rollouts without entry bias. Their number matches the number generated during latent-entry exploration in our method. Successful additional rollouts are retained under the same criterion. This control scores 3.8 points lower overall, with decreases across all three subsets, suggesting benefits beyond additional standard sampling.

Generalization to Single-Image Settings. Table 2 demonstrates the effectiveness of our method beyond multi-image settings. Our method achieves average scores of 47.7 on SPAR-Bench-SI and

75.8 on CV-Bench, improving over the base model by 23.8 and 5.1 points, respectively. It also surpasses HATCH by 2.5 and 1.3 points, achieving the highest average performance among the compared Qwen2.5-VL-3B-based models on both benchmarks. The gains across spatial reasoning in SPAR-Bench-SI and 2D/3D visual understanding in CV-Bench suggest that our method learns spatial representations that also benefit reasoning from a single view.

4 Related Work

4.1 Multi-Image Spatial Reasoning

Multi-image spatial reasoning requires models to infer spatial relationships from evidence distributed across multiple views (Yin et al., 2025; Yang et al., 2026a; Zhang et al., 2026a), where current MLLMs struggle to maintain a consistent understanding of the scene. Existing efforts broadly follow three directions. The first incorporates geometric priors from 3D foundation models such as VGGT (Wang et al., 2025a), either by injecting geometric features as additional inputs (Li et al., 2026c; Wu et al., 2026; Zhang et al., 2026b;c; Fan et al., 2026) or by distilling them into visual representations (Huang et al., 2026; Tang et al., 2026; Wang & Huang, 2026). These approaches primarily strengthen visual representations, but offer limited guidance on how geometric cues are integrated into a coherent scene representation. The second makes modeling explicit through intermediate supervision such as cognitive maps (Deng et al., 2026; Choi et al., 2026; Yin et al., 2025) and viewpoint-transition actions (Oi et al., 2026). The third enables models to construct and explore 3D scenes using external tools for reconstruction and novel-view rendering (Zhang et al., 2026d; Luo et al., 2026). Despite their effectiveness, the latter two rely on densely annotated supervision or external tools, which respectively limit scalability and increase inference overhead. Our work grounds spatial mental modeling in latent states through compact geometry tokens, without relying on dense spatial annotations or inference-time tool calls.

4.2 Latent Visual Reasoning

Latent reasoning in large language models (LLMs) performs reasoning through continuous hidden states rather than directly in discrete token space (Hao et al., 2024; Shen et al., 2025; Deng et al., 2025; Cheng & Van Durme, 2024; Wei et al., 2026; Tan et al., 2026). Recent work extends this idea to MLLMs as latent visual reasoning, interleaving visually grounded latent states with text generation to enable visual thinking during reasoning (Zhang et al., 2025; Yang et al., 2026b; Liu et al., 2026; Wang et al., 2026c; Jeon et al., 2026; Yu et al., 2026). LVR (Li et al., 2026a) supervises latent states to approximate region-of-interest features of the input image, while SLVR (Xu et al., 2026) further introduces attribute-level semantic supervision. In the spatial domain, recent efforts ground latent states in geometric information beyond what 2D visual embeddings provide. For single-image spatial reasoning, Aurora (Bigverdi et al., 2025) introduces perception tokens encoding depth maps and bounding boxes, while GeoAnchor (Li et al., 2026b) decomposes latent states into position, direction, and global geometry. Both require task-specific annotations as supervision, limiting scalability. For multi-image spatial reasoning, 3DThinker (Chen et al., 2026) aligns latent states with dense geometric features and enters latent mode at a fixed position before text generation. Our work revisits two design choices of this paradigm by aligning latent states with compact geometry tokens and learning flexible entry through reinforcement learning.

5 Conclusion

We introduce **Thinking with Latent Geometry Tokens**, a two-stage framework that equips MLLMs with spatial mental modeling in latent space for multi-image spatial reasoning. The first stage grounds latent states in compact camera and scene tokens from VGGT-Ω through supervised finetuning. We then augment GRPO with latent-mode entry exploration, using successful exploratory trajectories to guide policy updates. The resulting model reasons with geometry-grounded latent states without requiring VGGT-Ω at inference. Experiments show consistent gains on multi-image benchmarks, with improvements extending to single-image settings and across model families and scales. Ablations support the complementary roles of camera and scene tokens and the benefits of entry exploration beyond additional standard sampling. We hope this work advances spatial mental modeling in MLLMs and inspires future research on geometry-grounded latent reasoning.

AI USE STATEMENT

LLMs were used exclusively to assist with writing and editing, with their role restricted to improving grammar, spelling, clarity, and stylistic consistency. LLMs were also used for data curation. They were not involved in developing the scientific content, including formulating research ideas, designing experiments, analyzing data, interpreting results, or making other core intellectual contributions. All substantive content and scientific contributions originated entirely from the authors.

REPRODUCIBILITY STATEMENT

Our experiments use publicly available models (Qwen2.5-VL, VGGT-Ω) and datasets (MindCube-Train, SPAR-7M, and 3DThinker-10K). We describe the training data, model implementation, optimization settings, and evaluation procedures in the main text and appendix. Upon acceptance, we will release our model checkpoints and training data, along with the training and evaluation code, to support reproducibility and further research.

REFERENCES

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

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. *arXiv preprint arXiv:2502.13923*, 2025.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3836–3845. IEEE, 2025.

Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Personalizing multimodal large language models for image captioning: an experimental analysis. In *European conference on computer vision*, pp. 351–368. Springer, 2024.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024a.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2624, 2026.

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

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

Jiho Choi, Seonho Lee, Seojeong Park, and Hyunjung Shim. Dense reward for multi-view 3d reasoning with global maps and local views. *arXiv preprint arXiv:2606.23557*, 2026.

Jingcheng Deng, Liang Pang, Zihao Wei, Shicheng Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. Llm latent reasoning as chain of superposition. *arXiv preprint arXiv:2510.15522*, 2025.

Wei Deng, Xianlin Zhang, and Mengshi Qi. Active exploring like a pigeon: Reinforcing spatial reasoning via agentic vision-language models. *arXiv preprint arXiv:2606.02459*, 2026.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Google. Introducing Gemini 3. `https://blog.google/products-and-platforms/products/gemini/gemini-3-collection/`, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

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, 2026.

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

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *arXiv preprint arXiv:2602.04476*, 2026.

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

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026a.

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*, 2024a.

Hao Li, Han Fang, Zixin Pan, Xin Wei, Hongbo Sun, Jinglin Xu, Zhiyu Lin, Ye Yuan, Zhongjiang He, Yu Yu, et al. Geoanchor: Collaborative reasoning via latent decomposition for 3d spatial understanding. *arXiv preprint arXiv:2607.13454*, 2026b.

Haoyuan Li, Qihang Cao, Tao Tang, Kun Xiang, Zihan Guo, Jianhua Han, JiaWang Bian, Hang Xu, and Xiaodan Liang. Thinking with geometry: Active geometry integration for spatial reasoning. *arXiv preprint arXiv:2602.06037*, 2026c.

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, 2026d.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18061–18070. IEEE, 2024b.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9225–9236, 2026.

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

Masanari Oi, Koki Maeda, Ryuto Koike, Daisuke Oba, Nakamasa Inoue, and Naoaki Okazaki. From correspondence to actions: Human-like multi-image spatial reasoning in multi-modal large language models. *arXiv preprint arXiv:2602.08735*, 2026.

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

Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. Navbench: Probing multimodal large language models for embodied navigation. *Advances in Neural Information Processing Systems*, 38:95365–95385, 2026.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

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

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *Advances in Neural Information Processing Systems*, 38:4646–4668, 2026.

Jiayu Tang, Yuchen Zhou, and Chao Gou. Spatialsv: Internalizing interpretable 3d spatial awareness in mllms via task-oriented visual supervision. *arXiv preprint arXiv:2606.19915*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Haibo Wang and Lifu Huang. Learning geometric representations from videos for spatial intelligent multimodal large language models. *arXiv preprint arXiv:2606.05833*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-omega. *arXiv preprint arXiv:2605.15195*, 2026a.

Kangrui Wang, Linjie Li, Zhengyuan Yang, Shiqi Chen, Zihan Wang, Li Fei-Fei, Jiajun Wu, Leonidas Guibas, Lijuan Wang, and Manling Li. Planning with the views. *arXiv preprint arXiv:2605.29563*, 2026b.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026c.

Wei Wang, Yuqian Yuan, Tianwei Lin, Wenqiao Zhang, Siliang Tang, Jun Xiao, and Yueting Zhuang. Crossview suite: Harnessing cross-view spatial intelligence of mllms with dataset, model and benchmark. *arXiv preprint arXiv:2605.18621*, 2026d.

Yipu Wang, Yuheng Ji, Yuyang Liu, Enshen Zhou, Ziqiang Yang, Yuxuan Tian, Ziheng Qin, Yue Liu, Huajie Tan, Cheng Chi, et al. Towards cross-view point correspondence in vision-language models. *arXiv preprint arXiv:2512.04686*, 2025b.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 56721–56742, 2026.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 2026.

Tianrun Xu, Yue Sun, Qixun Wang, Jingyi Lu, Yuan Wang, Tianren Zhang, Longteng Guo, Fengyun Rao, Jing Lyu, Feng Chen, et al. Semantic-enriched latent visual reasoning. *arXiv preprint arXiv:2605.19342*, 2026.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026a.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33510–33520, 2026b.

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

Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31544–31555, 2026.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*, 2025.

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

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. *arXiv preprint arXiv:2603.27437*, 2026b.

Shihua Zhang, Qiuhong Shen, Shizun Wang, Tianbo Pan, and Xinchao Wang. Make geometry matter for spatial reasoning. *arXiv preprint arXiv:2603.26639*, 2026c.

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

A Evaluation Benchmarks

We evaluate our method on five evaluation sets from four benchmark families, covering multi-image and single-image spatial reasoning.

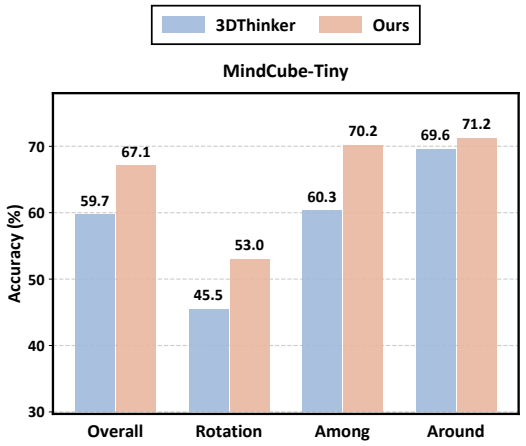


Figure 7: Comparison with 3DThinker on MindCube-Tiny using same geometric teacher model VGGT-Ω. Both methods use Qwen2.5-VL-3B as the backbone and are trained on 3DThinker-10K using VGGT-Ω for geometric supervision.

A.1 Multi-image Spatial Reasoning

SPAR-Bench-MV. SPAR-Bench-MV is the multi-view subset of SPAR-Bench (Zhang et al., 2026a), containing 4,345 questions with two or three images per question. It evaluates cross-view spatial relations, depth and distance estimation, and camera motion and viewpoint reasoning. We use accuracy for multiple-choice questions and mean relative accuracy (MRA) for numerical questions, and report difficulty-level and overall scores following the evaluation protocol of HATCH (Oi et al., 2026).

MindCube-Tiny. MindCube-Tiny is a compact subset of MindCube (Yin et al., 2025), containing 1,050 multiple-choice questions designed to assess spatial mental modeling from limited visual observations. It covers relative spatial relations, perspective transformations, and hypothetical observer movements. We report accuracy for the Rotation, Among, and Around categories, along with overall accuracy across all questions.

MMSI-Bench. MMSI-Bench (Yang et al., 2026a) contains 1,000 expert-designed multiple-choice questions on spatial intelligence in real-world multi-image scenes. It evaluates spatial relationships among cameras, objects, and regions, as well as attributes, motion, and multi-step reasoning. We report accuracy for Position, Attribute, Motion, and Multi-step Reasoning, together with overall accuracy across all questions.

A.2 Single-image Spatial Reasoning

SPAR-Bench-SI. SPAR-Bench-SI is the single-image subset of SPAR-Bench (Zhang et al., 2026a), containing 2,866 questions. It assesses depth and distance estimation, object spatial relations, and spatial imagination from a single image. We use accuracy for multiple-choice questions and MRA for numerical questions, and report difficulty-level and overall scores following the evaluation protocol of HATCH (Oi et al., 2026).

CV-Bench. CV-Bench (Tong et al., 2024) contains 2,638 single-image questions derived from established computer vision datasets. It evaluates 2D understanding through object counting and spatial relationships, and 3D understanding through depth ordering and relative distance. Following the official scoring protocol, we compute the 2D score by averaging the ADE20K and COCO source accuracies and the 3D score on Omni3D. The overall score is the mean of the 2D and 3D scores.



Figure 8: Additional qualitative comparisons with Qwen2.5-VL-3B. We show predictions and attention maps for both models, with attention differences in the rightmost column. Yellow boxes mark the target objects.

B Implementation Details

Baseline Evaluation. For all open-weight baselines in Table 1, we report results obtained using their publicly released checkpoints. In particular, the main results for 3DThinker (Chen et al., 2026) in Table 1 are obtained by evaluating the checkpoint released by its authors. Since only the SFT checkpoint of 3DThinker based on Qwen2.5-VL-3B (Bai et al., 2025) is publicly available, we use the results reported in the original paper for comparisons across model families and scales as shown in Figure 5.

C Additional Experiments

Controlling for the Geometric Teacher Model. To assess whether our gains stem from the choice of geometric teacher model, we retrain 3DThinker (Chen et al., 2026) using VGGT-Ω (Wang et al., 2026a) features for supervision, following its official default SFT configuration. Both methods use Qwen2.5-VL-3B as the backbone and are trained on 3DThinker-10K (Chen et al., 2026) through supervised finetuning. As shown in Figure 7, our method achieves 67.1% accuracy on MindCube-Tiny (Yin et al., 2025), compared with 59.7% for 3DThinker, a gain of 7.4 percentage points. Improvements hold across all three subsets, suggesting that our advantage is not solely attributable to the choice of geometric teacher.

D More Qualitative Results

Figure 8 presents two additional examples of viewpoint-dependent spatial reasoning. In the first case, our method correctly identifies the printed glass door behind the reference viewpoint, while the base model selects the window. In the second case, our method correctly locates the TV to the right of the white square stool with pillows, whereas the base model selects a leather loveseat. The attention visualizations show greater focus on the target regions in both cases, suggesting that our method may better identify relevant visual cues across views.