ARTEMIS-3D: LATENT SPATIAL REASONING FOR VISUAL-SPATIAL INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

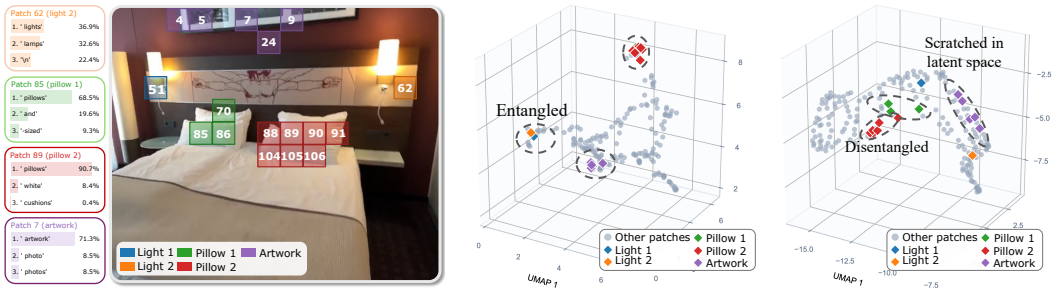
Visual-spatial intelligence poses significant challenges for current multimodal large language models (MLLMs), whose language-centric representations limit their ability to reason about spatial relationships. Existing approaches address this limitation by either injecting geometric information through external modules or verbalizing spatial information into textual Chain-of-Thought (CoT), yet both ultimately retain a language-oriented reasoning space. Meanwhile, we observe that language-oriented visual representations can conflate semantically similar objects at distinct spatial locations, obscuring the spatial distinctions essential for reasoning. To overcome the constraints of language-oriented spatial reasoning, we introduce *Latent Spatial Reasoning*, a new paradigm that enables MLLMs to construct and reason directly within a spatially grounded latent space. We instantiate this paradigm in **Artemis-3D**, a unified framework with two complementary components. First, *Mental Map Imaging (MMI)* internalizes geometric structure into visual token representations, constructing a latent mental map that preserves spatial distinctions. Second, *Visual-Cue Latent Reasoning (VCLR)* autoregressively retrieves question-relevant spatial evidence from this mental map and performs variable-length reasoning directly in the spatial latent space. Together, MMI and VCLR enable Artemis-3D to preserve geometric structure throughout the reasoning process without explicit textual CoT. Extensive experiments demonstrate that Artemis-3D outperforms state-of-the-art methods on visual-spatial intelligence benchmarks. Our code will be publicly available.

1 INTRODUCTION

Visual-spatial intelligence (VSI) (Yang et al., 2025a) is a challenging problem that extends beyond conventional 2D visual understanding, requiring models to perceive geometric structure and reason about complex spatial relationships. Despite their strong capabilities in visual understanding and reasoning, current multimodal large language models (MLLMs) face significant challenges in these tasks. Early attempts (Wu et al., 2025; Zheng et al., 2025; Fan et al., 2025) enhance spatial understanding by incorporating explicit 3D representations, such as point clouds, or implicit geometric features extracted from geometric foundation models (Wang et al., 2025b; Lin et al., 2025; Wang et al., 2024) through plug-in modules. More recent spatial reasoning methods (Gao et al., 2026; Chen et al., 2025c; Yin et al., 2025; Chen et al., 2025b) construct textual Chain-of-Thought (CoT) by verbalizing spatial information, such as bird's-eye-view maps or 3D object positions, as intermediate reasoning traces before producing the final answer. However, despite these advances, existing methods still lack an internal representation space specifically suited for spatial reasoning. Geometry-enhanced approaches primarily inject spatial information from external modules, while spatial reasoning methods still express their intermediate reasoning process in language. As a result, spatial reasoning is ultimately constrained to a language-oriented representation space, which is inherently mismatched with the geometric nature of VSI.

In this work, we propose to perform reasoning directly within the MLLM's internal spatial latent space, rather than verbalizing spatial information into language as intermediate reasoning traces. We term this new paradigm ***Latent Spatial Reasoning***. Instead of treating external 3D representations as additional reasoning inputs, our approach encourages the MLLM to internalize geometric structure within its own latent states and reason directly over these spatially grounded representations. This

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.



(a): De-tokenized visual tokens from a MLLM (b): Visual tokens of a MLLM (c): Visual tokens of VGGT

Figure 1: The collapse of visual representations into the language space. (a) De-tokenized output visual tokens from a MLLM exhibit clear semantic correspondence to the image patches. (b) Visualization of these visual tokens shows that tokens from the same semantic category cluster closely despite distinct spatial locations. (c) In contrast, visual tokens from a geometric foundation model remain well separated across different spatial locations.

enables implicit spatial reasoning without relying on explicit textual intermediate steps, while preserving spatial information throughout the reasoning process.

Our approach is motivated by two key observations. First, current MLLMs lack an internal representation mechanism explicitly designed to preserve the geometric structure of a visual scene. Existing geometry-enhanced methods typically introduce 3D information through external representations or auxiliary modules, meaning that geometric structure is supplied to the model rather than internally established within the LLM itself. Second, we observe that language-oriented visual representations can conflate semantically similar objects at distinct spatial locations. As illustrated in Fig. 1(a), we detokenize the visual tokens from an MLLM (e.g., Qwen3-VL (Bai et al., 2025a)) using its LM head. The resulting vocabulary distributions closely reflect the semantic content of individual patches, such as "lights" and "pillows", indicating strong semantic alignment. However, the UMAP (McInnes et al., 2018) visualization in Fig. 1(b) reveals that tokens representing spatially separated objects of the same category (e.g., lights or pillows) form tight clusters, suggesting that these representations preserve semantic similarity while obscuring spatial distinctions. In contrast, the last-layer representations of a geometric foundation model (e.g., VGGT (Wang et al., 2025b)) separate these object instances more clearly in Fig. 1(c). These qualitative comparisons suggest that the MLLM's visual representations emphasize the semantics of the objects, while providing less spatial separation between distinct instances of the same semantic category. This preference for semantic information is also consistent with prior findings (Brown et al., 2025; Huang et al., 2025a; Ma et al., 2026; Min et al., 2026) that MLLMs can take semantic shortcuts when answering spatial questions, relying on semantic associations rather than the underlying 3D geometry. Our analysis extends these behavioral observations to the representation level: the preference for semantic similarity over spatial distinction is already apparent in visual-token states within the MLLM's latent space.

Taken together, these observations reveal that supplying additional geometric information or adding textual CoT to an MLLM is inherently a limited and indirect way for improving spatial reasoning. On the contrary, the proposed *Latent Spatial Reasoning* offers a more natural way to enhance spatial reasoning by enabling the model to construct and reason directly within a spatially grounded latent space. We instantiate this insight in **Artemis-3D**, a framework consisting of two complementary components: 1) *Mental Map Imaging (MMI)*, which enables the MLLM to internally generate *geometric tokens* that encode the 3D structure of the scene as a latent mental map, and 2) *Visual-Cue Latent Reasoning (VCLR)*, which leverages these geometric tokens to autoregressively generate *spatial latent tokens* and reason directly within this internal latent space.

Specifically, *Mental Map Imaging* imposes geometric supervision on the hidden states of visual tokens at intermediate LLM layers and introduces dedicated camera and scale tokens together with learnable scene tokens that aggregate global information across views. Depth, point-map, camera, and scale reconstruction jointly supervise this internal representation. In this way, geometric knowledge is internalized within the model's own latent space. The resulting geometric tokens encode the global 3D structure of the scene and collectively form a latent mental map, which can optionally be decoded into explicit 3D predictions for interpretation. Importantly, the explicit geometric supervision and reconstruction objectives are confined to the training stage. At inference time, Artemis-3D therefore

constructs its mental map directly from its internal representations without relying on external geometric foundation models or additional 3D feature extraction modules.

Building upon this latent mental map, *Visual-Cue Latent Reasoning* enables the model to autoregressively generate *spatial latent tokens* that retrieve spatial evidence from the geometric tokens in the mental map. Each retrieved geometric token is fed back into the model to condition the next reasoning step, forming a reasoning trace that progressively accumulates question-relevant spatial evidence while preserving its geometric structure. During training, spatial latent tokens are learned through a next-latent embedding prediction objective with cosine similarity loss, while an additional progress head predicts when the latent reasoning stage should terminate. At inference time, each generated spatial latent token is matched to the most similar geometric token in the mental map, which is then fed back to initiate the next reasoning step. This process continues autoregressively until the progress head signals completion, after which language generation is invoked to produce the final answer.

Experiments on three representative visual-spatial intelligence benchmarks demonstrate the effectiveness of Artemis-3D. On VSI-Bench, our 2B model achieves 72.5% mean accuracy, outperforming the larger Cambrian-P-7B (Yang et al., 2026). Beyond camera pose estimation of Cambrian-P, our method also predicts dense depth and point maps, enabling comprehensive spatial mental modeling. On MindCube (Yin et al., 2025), our method outperforms SpatialDreamer (Cao et al., 2026) by 2.3%, despite the latter using an external world model to generate imagined views. Moreover, it outperforms ChatScene (Huang et al., 2023a) by 19.9% on Real-3DQA (Ma et al., 2026) without requiring 3D point cloud inputs. Further analyses and visualizations show that MMI's geometirc tokens retain reconstructable geometry, while VCLR generates variable-length reasoning traces in a latent space distinct from language representations.

Our contributions can be summarized as follows:

- We introduce *Latent Spatial Reasoning*, a new paradigm that enables MLLMs to construct spatially grounded latent representations and reason directly within them, rather than verbalizing spatial reasoning into textual CoT.
- We propose Artemis-3D, a unified framework combining *Mental Map Imaging* for constructing an internal geometric representation and *Visual-Cue Latent Reasoning* for autoregressively reasoning to gather question-relevant evidence in the spatial latent space.
- Extensive experiments and visualizations demonstrate that Artemis-3D reconstructs 3D structures and generates variable-length latent spatial reasoning trace, outperforming state-of-the-art methods across multiple visual-spatial intelligence benchmarks.

2 Related Work

Multimodal Large Language Models for 3D Understanding. Existing methods enhance 3D understanding in MLLMs through geometric inputs or geometry-aware supervision. The former encode scenes using object proposals (Huang et al., 2023b; Zhu et al., 2023; Kang et al., 2024), aggregated multi-view features (Hong et al., 2023; Zhu et al., 2024), or dedicated 3D encoders (Chen et al., 2024; Deng et al., 2025). More recently, VG LLM (Zheng et al., 2025) augments visual tokens with features from geometric foundation models (e.g., VGGT (Wang et al., 2025b)), while related methods explore visual–geometric fusion and feature injection into the language backbone (Li et al., 2026; Zhang et al., 2026b; Liu et al., 2026; Zhang et al., 2026c; Wang et al., 2026a; Zhang et al., 2026a; Chen et al., 2025a; Liang et al., 2025; Meng et al., 2024). Another line of work, including MILO (Cao et al., 2025), ROSS3D (Wang et al., 2025a), GeoVR (Wang et al., 2026b), SpatioLM (Wu et al., 2026a), and S^2-MLLM (Xu et al., 2025), learns spatial representations through geometric supervision or reconstruction objectives and uses 3D features to augment the representations fed into MLLMs, yet the subsequent reasoning remains language-oriented. In contrast, we use geometric foundation models only to generate pseudo-labels, from which the model learns to construct its own internal mental map. More importantly, this mental map is not designed merely to reconstruct the 3D scene; it establishes a spatially grounded latent space that enables the subsequent VCLR to perform native spatial reasoning.

Latent Reasoning. Latent reasoning (Yu et al., 2026) initially emerged in the NLP community as a means of replacing intermediate textual reasoning steps with continuous hidden states. CO-

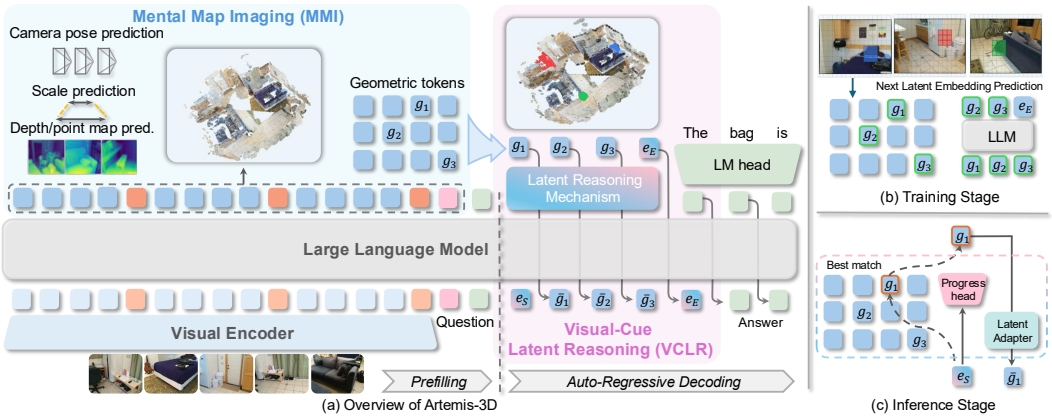


Figure 2: Overview of Artemis-3D. Mental Map Imaging (MMI) constructs an internal spatial representation of the scene within the MLLM's hidden states. Based on this representation, Visual-Cue Latent Reasoning (VCLR) performs reasoning in the spatial latent space, generating a token-level visual reasoning trace that captures the crucial visual cues.

CONUT (Hao et al., 2024) advances this paradigm, with subsequent work exploring soft concept tokens (Zhang et al., 2025), continuous chain-of-thought (Shen et al., 2025), and compressed representations (Cheng & Durme, 2024). More recently, this paradigm has been extended to the vision-language domain by incorporating visual evidence into latent reasoning. MCOUT (Pham & Ngo, 2025) refines joint visual–textual states, while LVR (Li et al., 2025a) reconstructs relevant visual tokens and LIVR (Li et al., 2025b) learns visual reasoning tokens without intermediate supervision. CoVT (Qin et al., 2025) encodes dense perceptual cues, Monet (Wang et al., 2025c) learns visual embeddings through distillation and policy optimization, and DMLR (Liu et al., 2025) combines test-time latent optimization with dynamic visual injection. Compared with prior visual latent reasoning methods (Li et al., 2025a; Qin et al., 2025; Li et al., 2025b), our approach differs in two respects: 1) it extends latent reasoning to the *spatial domain* by operating on geometric tokens from the MLLM's internal mental map; and 2) VCLR generates stable, variable-length latent spatial reasoning trajectories that can span hundreds of steps.

3 Method

Our framework is designed to enable MLLMs to both preserve geometric structure within their internal representations and reason directly over such structure without translating it into language. As shown in Fig. 2, Artemis-3D consists of two complementary components: 1) *Mental Map Imaging (MMI)* that enables the MLLM to internalize scene geometry within its visual-token representations and construct a latent mental map and 2) *Visual-Cue Latent Reasoning (VCLR)* that performs reasoning directly in the spatial latent space by autoregressively generating spatial latent tokens and retrieving task-relevant geometric evidence from the mental map. We describe each component in detail.

3.1 Mental Map Imaging

Given N multi-view observations and a question, the vision encoder maps each view to visual tokens $\mathbf{V}_i$. We append a camera token $\mathbf{c}_i$ to each view, followed by M learnable scene tokens $\mathbf{U} = [\mathbf{u}_1; \ldots; \mathbf{u}_M]$ and a scale token $\mathbf{s}$ to form the augmented visual sequence $\widetilde{\mathbf{V}}$:

$$\widetilde{\mathbf{V}} = [\mathbf{V}_1; \mathbf{c}_1; \ldots; \mathbf{V}_N; \mathbf{c}_N; \mathbf{U}; \mathbf{s}]. \tag{1}$$

In order to internalize a spatial mental model within the MLLM, we supervise $\widetilde{\mathbf{V}}$ with geometric reconstruction objectives, including depth, point maps, and camera parameters. We utilize frozen geometry foundation models to generate pseudo-labels for these objectives. We use separate DPT heads (Ranftl et al., 2021) to fuse multi-level visual features and predict dense depth and point maps, respectively. At each layer selected by the DPT heads, visual tokens attend to the scene tokens $\mathbf{U}$ through gated cross-attention, enabling local geometric predictions to incorporate global scene context. A cross-frame camera decoder jointly processes the camera-token states and iteratively

refines the translation, quaternion rotation, and field of view. In parallel, a lightweight MLP predicts the log metric scale from the scale-token state.

The depth objective $\mathcal{L}_{\text{depth}}$ uses an ℓ_1 loss, while the point objective $\mathcal{L}_{\text{point}}$ adopts a confidence-weighted smooth ℓ_1 loss. The camera objective $\mathcal{L}_{\text{cam}}$ combines weighted ℓ_1 losses over translation, quaternion rotation, and field of view, and the metric scale is supervised with a smooth ℓ_1 loss. The overall MMI objective is

$$\mathcal{L}_{\text{MMI}} = \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{point}}\mathcal{L}_{\text{point}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}}. \tag{2}$$

These reconstruction objectives internalize geometric knowledge within the MLLM $\mathcal{F}_\theta$. We extract the final hidden states at visual-token positions $\mathcal{V}$ as the *geometric token bank*:

$$\mathbf{G} = \{g_j\}_{j=1}^{|\mathcal{V}|} = \text{stopgrad}\Big(\mathcal{F}_\theta([\widetilde{\mathbf{V}}; \mathbf{T}_q])\big|_{\mathcal{V}}\Big). \tag{3}$$

The geometric token bank $\mathbf{G}$ retains patch-level geometric evidence and serves as the candidate bank for subsequent visual-cue token retrieval in VCLR.

3.2 Visual-Cue Latent Reasoning

3.2.1 Visual-Cue Trace Preparation

As shown in Fig. 3, answering spatial questions requires attending to question-relevant visual regions and reasoning about the spatial relationships between them. We aim to realize this reasoning process in the spatial latent space. To support this process, we construct a visual-cue reasoning trace from geometric tokens associated with these regions. This reasoning trace guides the model through spatially grounded regional evidence before answer generation.

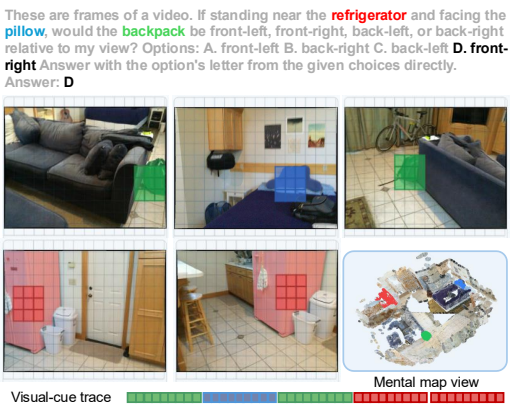


Figure 3: Question-relevant visual regions and their visual patches across video frames.

Specifically, we develop an automatic annotation pipeline that extracts question-relevant visual regions as bounding boxes in the input images. We map each retained box through the image preprocessing transformations onto the visual encoder's token grid and select an $n \times n$ neighborhood around its center. These neighborhoods provide patch-level representations of the cue regions $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_L)$. The union of the selected positions across retained regions defines the candidate positions for the trace. We arrange these positions into a sequence of indices and retrieve the corresponding states from the geometric token bank $\mathbf{G}$ to form the target visual-cue trace $\mathbf{L} = (g_1, \ldots, g_K)$. We insert this reasoning trace between the question and the textual answer, bracketed by special embeddings that mark the start and end of latent spatial reasoning. The resulting LLM input sequence is

$$\mathcal{X} = [\widetilde{\mathbf{V}}; \mathbf{T}_q; \mathbf{e}_S; \mathbf{L}; \mathbf{e}_E; \mathbf{T}_y], \tag{4}$$

where $\mathbf{e}_S$ and $\mathbf{e}_E$ denote the start and end embeddings, respectively, and $\mathbf{T}_y$ denotes the textual answer tokens. This construction integrates region-level geometric evidence into the LLM's standard left-to-right computation, allowing latent spatial reasoning over visual cues to transition directly into textual answer generation.

3.2.2 Reasoning in the Spatial Latent Space

Next-latent embedding prediction. Our goal is to enable the model to autoregressively generate the visual-cue trace $\mathbf{L}$ within the spatial latent space. Inspired by the autoregressive principle of standard language modeling, we replace next-token prediction over a discrete vocabulary with next-latent embedding prediction. Specifically, VCLR predicts each visual cue conditioned on the adapted geometric tokens from preceding reasoning steps:

$$\widehat{g}_k = \mathcal{F}_\theta\Big(\widetilde{\mathbf{V}}, q, \mathcal{A}(g_{<k})\Big), \qquad k = 1, \ldots, K, \tag{5}$$

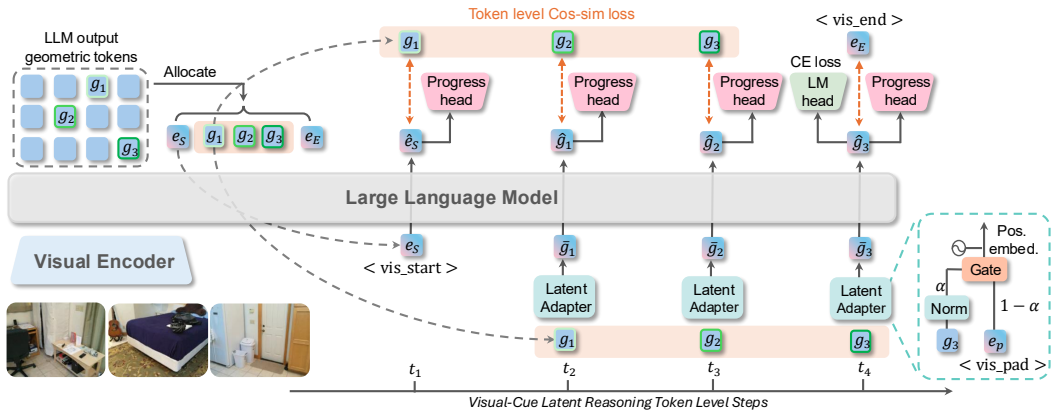


Figure 4: Training of Visual-Cue Latent Reasoning. Geometric tokens provide supervision for next hidden embedding prediction and are fed through a latent adapter to condition subsequent reasoning steps. A progress head learns when the latent reasoning trajectory should terminate.

where $\mathcal{A}$ denotes the latent adapter that maps geometric tokens into the LLM input embedding space, and $\mathcal{A}(g_{<k}) = \{\mathcal{A}(g_1), \ldots, \mathcal{A}(g_{k-1})\}$. Unlike conventional next-token prediction, each reasoning step directly predicts a continuous spatial latent embedding rather than a probability distribution over discrete vocabulary tokens.

Latent token supervision. As shown in Fig. 4, given the preceding visual cues, the LLM output hidden state $\widehat{g}_k$ at step k predicts the next hidden embedding. We optimize this prediction toward the corresponding target geometric token g_{k+1} using a cosine-similarity objective:

$$\mathcal{L}_{\text{NLEP}} = \frac{1}{K} \sum_{k=0}^{K-1} \left[1 - \cos(\widehat{g}_k, g_{k+1})\right]. \tag{6}$$

During training, each target geometric token is passed through the latent adapter and fed into the LLM to condition the prediction of the subsequent cue, forming an autoregressive reasoning trace entirely within the spatial latent space.

3.2.3 Stablizing the spatial latent reasoning.

Directly predicting the next hidden embedding in the spatial latent space can be unstable (Hao et al., 2024) and existing methods usually using fixed latent reasoning steps (Li et al., 2025a; Qin et al., 2025; Li et al., 2025b). To address this, we introduce a latent adapter that maps geometric tokens into the LLM input space, mitigating distribution mismatch and stabilizing training and inference. Additionally, we incorporate a progress head that predicts when the latent reasoning stage should terminate, allowing for variable-length reasoning traces that adapt to the complexity of the spatial question.

Latent adapter. The latent adapter $\mathcal{A}$ maps each geometric token g_k into the LLM input space through gated interpolation:

$$\widetilde{g}_k = (1 - \sigma(\alpha))\mathbf{e}_{\text{pad}} + \sigma(\alpha)\, P(\text{Norm}(g_k)), \tag{7}$$

where P is a linear projection, $\mathbf{e}_{\text{pad}}$ is a latent placeholder embedding. Since M-RoPE does not distinguish latent steps at the same template position, we add a gated sinusoidal step embedding:

$$\bar{g}_k = \widetilde{g}_k + \sigma(\beta)\ \text{RMSMatch}\big(\text{PE}_{\sin/\cos}(k), \widetilde{g}_k\big), \tag{8}$$

where RMSMatch matches the step embedding's magnitude to that of $\widetilde{g}_k$. The resulting embedding $\bar{g}_k$ conditions the next visual-cue prediction.

Progress head. For each latent reasoning step k, a simple MLP progress head predicts a distribution $\text{P}_k(r)$ over the number of remaining cue regions, where $r \in \{\varnothing, \ldots, L\}$. The class $\varnothing$ indicates completion and $\text{P}_k(\varnothing)$ defines the binary completion probability. A region is resolved once all its selected visual tokens have been covered. The progress objective $\mathcal{L}_{\text{prog}}$ combines region-weighted cross-entropy, an ℓ_1 loss on the expected remaining count, and a penalty for increases in this count

across steps. Moreover, we use a balanced binary cross-entropy loss $\mathcal{L}_{\text{stop}}$ to supervise the stopping signal. For a complete trace, steps $1, \ldots, K-1$ are supervised to continue, whereas the last step K is supervised to stop:

$$\mathcal{L}_{\text{stop}} = -\frac{1}{2(K-1)} \sum_{k=1}^{K-1} \log(1 - \mathrm{P}_k(\varnothing)) - \frac{1}{2} \log \mathrm{P}_K(\varnothing). \tag{9}$$

At inference, latent reasoning stops once $\mathrm{P}_K(\varnothing)$ reaches a threshold τ, enabling variable reasoning lengths through adaptive termination.

3.2.4 VCLR Inference

At inference stage, VCLR proceeds autoregressively by iteratively retrieving relevant geometric tokens and feeding them back into the MLLM to guide subsequent latent predictions. At each step k, the predicted spatial latent token $\widehat{g}_k$ is matched against the geometric token bank $\mathbf{G}$ using cosine similarity. Previously selected bank indices are masked, and the highest-scoring geometric token is retrieved, so each index can be selected at most once. The retrieved token is passed through the latent adapter and fed back into the MLLM to condition the next visual-cue latent token prediction. Latent reasoning terminates when the progress head's completion probability $\mathrm{P}_k(\varnothing)$ reaches the threshold τ or reachs the maximum reasoning step length L_{max}. The model then consumes the visual-cue-end token $\mathbf{e}_E$ and resumes standard textual autoregressive decoding to generate the answer. Further inference details are provided in Appendix A.1.

4 Experiments

4.1 Implementation Details

We train the model for one epoch on the VSI-590K (Yang et al., 2025c) and VLM-3R (Fan et al., 2025) datasets. We use 32 uniformly sampled frames per scene. We use VGGT (Wang et al., 2025b) to generate pseudo-labels for depth, point maps, camera parameters, and use MapAnything (Keetha et al., 2025) for the metric scale. Detailed training settings and component configurations are provided in Appendix A.3. We use LMMs-Eval (Zhang et al., 2024) for benchmark evaluation.

4.2 Experimental Results

Spatial intelligence benchmarks. We evaluate our method on three representative spatial intelligence benchmarks: VSI-Bench (Yang et al., 2025a), MindCube-Tiny (Yin et al., 2025), and Real-3DQA (Ma et al., 2026). In Table 1, Artemis-3D-2B exceeds GeoVR-2B by 3.4%, and outperform Cambrian-P (Yang et al., 2026) with 7B model size. On MindCube-Tiny (Table 2a), it achieves 87.2% overall accuracy, surpassing SpatialSV-8B (Tang et al., 2026) by 13.5% and SpatialDreamer (Cao et al., 2026) by 2.3%. On Real-3DQA (Table 2b), Artemis-3D achieves the highest EM_R among the compared methods (Hong et al., 2023; Huang et al., 2023b; Wang et al., 2023; Huang et al., 2023a), outperforming GPT4Scene-HDM (Qi et al., 2026) by 5.1%. While GPT4Scene-HDM uses extra BEV maps and detected object markers from Mask3D (Schult et al., 2023), our method relies solely on multiview images as visual input. Together, these results demonstrate that our method achieves superior performance on spatial intelligence tasks with a smaller, more compact model, without requiring any additional 3D information.

Analysis of VCLR. Figure 5 examines the representation and reasoning behavior of VCLR. In (a), the UMAP visualization shows visual-cue latent tokens overlapping with visual embeddings while remaining separated from text embeddings, supporting the view that VCLR reasons in a spatial latent space distinct from the text latent space. In (b), visual-cue latent reasoning trace lengths vary from fewer than 20 to over 200 steps, demonstrating variable-length latent reasoning. In (c), stop decisions are associated with lower token-matching cosine similarity than continue decisions, suggesting that the progress head learns to terminate VCLR as alignment with visual evidence weakens. Overall, these observations support VCLR's ability to perform adaptive and controllable reasoning grounded in spatial representations.

Effect of MMI and VCLR. For the ablations, we train each model on the video subset of VSI-590K for one epoch and evaluate it on VSI-Bench. Table 3 shows that both MMI and VCLR contribute to

Table 1: Comparison with state-of-the-art models on VSI-Bench. † denotes the result of Cambrian-P trained with 32 input frames for a fair comparison.

Model	Size	Avg.	Numerical Answer				Multiple-Choice Answer			
			Obj. Co.	Ab. Dis.	Obj. Si.	Roo. Si.	Re. Dis.	Rel. Dir.	Rou. Pl.	App. Ord.
General-purpose Models										
GPT-4o (Hurst et al., 2024)	–	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-2.5 Pro (Comanici et al., 2025)	–	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
InternVL-3.5 8B (Wang et al., 2025d)	8B	56.3	–	–	–	–	–	–	–	–
Qwen2.5-VL-7B (Bai et al., 2025b)	7B	29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
Qwen3-VL-2B (Bai et al., 2025a)	2B	50.3	62.1	40.2	71.4	49.7	52.2	42.0	30.4	54.5
Qwen3-VL-8B (Bai et al., 2025a)	8B	57.9	67.5	47.0	76.3	61.9	58.0	50.9	35.0	66.3
Spatially Enhanced Models										
3DRS (Huang et al., 2025b)*(NeurIPS 25)*	7B	45.9	68.7	34.8	53.6	56.6	40.9	43.2	30.4	39.2
Spatial-MLLM (Wu et al., 2025)*(NeurIPS 25)*	4B	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VEGA-3D (Wu et al., 2026b)*(ECCV 26)*	8B	50.5	69.7	35.9	58.0	60.8	45.1	43.1	30.9	60.5
GeoThinker (Li et al., 2026)*(ICML 26)*	8B	50.5	69.5	38.5	57.9	62.2	45.2	46.2	31.4	52.6
VG LLM (Zheng et al., 2025)*(NeurIPS 25)*	8B	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
VLM-3R (Fan et al., 2025)*(CVPR 26)*	7B	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
VST-7B-SFT (Yang et al., 2025b)*(ECCV 26)*	7B	60.6	72.0	44.4	74.3	68.3	59.7	55.8	44.9	65.2
Map2Thought (Gao et al., 2026)*(CVPR 26)*	7B	61.0	70.8	55.0	70.1	69.4	56.9	69.8	38.1	57.4
GS-Reasoner (Chen et al., 2025b)*(ICLR 25)*	7B	64.7	69.1	61.9	70.0	65.7	65.4	88.9	44.3	52.3
Cambrian-S (Yang et al., 2025c)*(ICLR 26)*	7B	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
SenseNova-SI (Cai et al., 2025)*(CVPR 26)*	8B	68.7	-	-	-	-	-	-	-	-
GeoVR-2B (Wang et al., 2026b)*(arXiv 26)*	2B	69.1	67.7	54.5	73.9	72.3	71.3	80.7	45.9	86.7
GeoVR-4B (Wang et al., 2026b)*(arXiv 26)*	4B	69.6	65.6	58.1	75.5	71.6	70.0	89.3	41.2	85.6
SpaceMind (Zhao et al., 2025)*(CVPR 26)*	9B	69.6	73.3	61.4	77.3	74.2	67.2	88.4	44.3	70.6
SpatioLM (Wu et al., 2026a)*(ICML 26)*	8B	71.6	81.9	55.7	75.2	71.5	78.2	74.4	47.9	87.8
Cambrian-P† (Yang et al., 2026)*(ECCV 26)*	7B	72.0	–	–	–	–	–	–	–	–
Artemis-3D-2B (Ours)	**2B**	**72.5**	74.0	59.4	79.4	75.0	75.1	81.8	50.0	85.1
Artemis-3D-4B (Ours)	**4B**	**72.6**	73.7	61.5	79.0	76.8	72.3	86.0	46.4	85.0

Table 2: Results on (a) MindCube-Tiny and (b) Real-3DQA. In (a), † denotes Qwen2.5-VL-3B-Instruct under the MindCube SFT configurations. In (b), VRS is the Viewpoint-Rotation Score and baseline results are from Ma et al. (2026). * indicates that GPT4Scene uses extra BEV map and detected object markers as MLLM inputs.

Method	**MindCube-Tiny QA Accuracy (%)**			
	Rotation	Among	Around	Overall ↑
Qwen3-VL-2B	34.0	32.3	38.8	34.1
Qwen2.5-VL-3B†	47.5	62.3	67.6	60.8
3DThinker-SFT	44.0	64.8	72.4	62.7
SpatialSV-8B	38.5	77.2	86.0	73.7
SpatialDreamer	87.5	78.2	93.5	84.9
Artemis-3D-2B (Ours)	69.5	93.6	86.0	87.2

(a) MindCube-Tiny

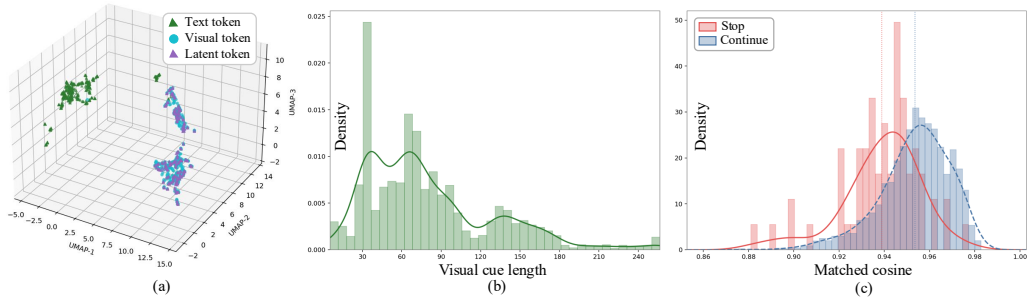
Method	**Real-3DQA**		**Correct times**				**VRS**
	EM	EM_R	one	two	three	four	%
3D-LLM	7.5	10.4	33.2	4.1	1.1	0.1	9.6
Chat-3D v2	3.4	9.7	23.2	2.7	0.5	0.0	6.6
LEO	14.3	19.1	46.9	8.1	1.6	0.4	14.3
Chat-Scene	17.0	22.1	43.3	7.1	1.2	0.1	12.9
GPT4Scene-HDM*	33.1	36.9	55.5	14.3	2.5	0.5	18.2
Artemis-3D (Ours)	36.6	42.0	41.5	10.5	1.7	0.3	13.5

(b) Real-3DQA

spatial reasoning performance. Starting from the Qwen3VL-2B SFT baseline, VCLR alone improves the mean accuracy by 0.8%, while MMI alone improves 2.0%. Combining MMI and VCLR yields a significant 4.5% improvement. These results show that both MMI and VCLR improve spatial reasoning individually, while their combination produces a significant synergistic effect, yielding a gain greater than the sum of their individual improvements.

Component Analysis of MMI and VCLR. In Table 4a, adding the latent adapter, grouped object traces, and progress head progressively improves VCLR by 2.7% mean accuracy, supporting structured latent reasoning with adaptive termination. For MMI (Table 4b), joint depth and point supervision achieves better results than using either alone, and scene tokens further improve the score by 1.3%. Table 4c shows the no-repeat VCLR decoding strategy performs best. Table 4d reports geometric reconstruction results on ScanNet-Val (Dai et al., 2017). Although a performance gap remains relative to the specialized reconstruction model VGGT, Artemis-3D demonstrates the potential for reconstructing 3D scene geometry solely from the MLLM's internal representations.

Visualizations. Figure 6 illustrates how Artemis-3D constructs an internal mental map of the scene and reasons over spatial cues to answer a visual-spatial question. The latent tokens predicted during VCLR are similar to the image patches (highlighted with cyan boxes) corresponding to cue objects referenced in the question (e.g., the chair, backpack, and door). Moreover, the geometric tokens can



(a) (b) (c)

Figure 5: (a) Visualization of output embeddings. (b) Distribution of VCLR visual-trace lengths. (c) Matching cosine-similarity distributions for stop and continue decisions by the progress head.

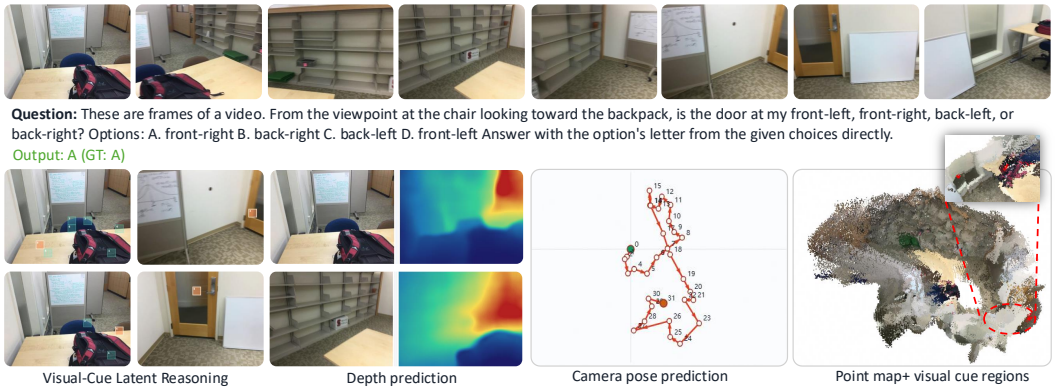


Figure 6: Visualization of Artemis-3D's latent spatial reasoning process for a visual-spatial question. Cyan boxes indicate matched visual tokens in VCLR, while red masks highlight their corresponding locations in the 3D point cloud reconstructed from MMI.

Table 3: Effect of the proposed MMI and VCLR. (0) denotes the Qwen3VL-2B SFT baseline.

#	MMI	VCLR	Avg.	Numerical Answer				Multiple-Choice Answer			
				Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
(0)	–	–	65.2	64.6	50.2	75.6	67.9	66.8	71.5	48.4	71.7
(1)	–	✓	66.0 (+0.8)	71.6	48.9	75.5	67.0	69.9	76.7	36.1	82.2
(2)	✓	–	67.2 (+2.0)	71.4	51.2	75.1	71.0	72.4	80.3	37.1	79.3
(3)	✓	✓	**69.7 (+4.5)**	72.6	53.6	78.1	71.9	70.1	80.7	44.9	85.8

Table 4: Ablation studies: (a) key components of VCLR; (b) key components of MMI; (c) decoding strategies in VCLR; and (d) point cloud reconstruction quality on ScanNet-Val.

Latent Adapter	Grouped Object Trace	Progress Head	Avg.
–	–	–	67.0
✓	–	–	68.4
✓	✓	–	69.3
✓	✓	✓	69.7

(a)

Depth Head	Point Head	Scene Tokens	Avg.
✓	–	–	67.4
–	✓	–	67.2
✓	✓	–	68.4
✓	✓	✓	69.7

(b)

Decoding Strategy	Avg.
Greedy	67.7
Matching	68.4
Beam Search	68.8
No Repeat	69.7

(c)

Method	Acc.↓	Comp.↓	F-1 ↑
VGGT	0.13	0.11	0.46
Artemis-3D			
Depth&Cam	0.28	0.37	0.21
Point	0.15	0.17	0.36

(d)

be decoded to recover approximate scene geometry, including depth, camera pose, and a scene-level point cloud. Additional qualitative visualizations are provided in Appendix A.4.

5 Conclusion

In this work, we introduced *Latent Spatial Reasoning*, a new paradigm that enables MLLMs to construct spatial representations and reason directly in a geometry-grounded latent space rather than through textual CoT. We instantiated this idea in Artemis-3D, which combines Mental Map Imaging for internalizing 3D scene structure and Visual-Cue Latent Reasoning for retrieving question-relevant spatial evidence. Experiments demonstrate strong performance across spatial intelligence benchmarks. We hope this work inspires further exploration of latent spatial reasoning for embodied AI and general-purpose spatial intelligence.

AI USE STATEMENT

We used large language models to assist with drafting and polishing manuscript prose, including sentence rephrasing, grammar correction, and improvements to clarity and flow. These writing assistants were not used to design the method, run experiments, or select the reported results. The authors reviewed the AI-assisted text against the underlying methods and experimental results, revised it where necessary, and take full responsibility for the content of the paper.

ETHICS STATEMENT

This work studies visual-spatial reasoning using existing third-party datasets and pretrained models. We did not recruit participants, collect new human-subject data, or conduct animal experiments. The training and evaluation data include real indoor scenes and videos from third-party sources, which may contain people, personal spaces, or other sensitive visual information. Our experiments use these data for scene understanding and spatial question answering; they do not seek to identify individuals, track people, or infer sensitive personal attributes. Researchers using or redistributing the underlying data should follow the access conditions and licenses of the original sources. Dataset coverage may also favor particular environments, viewpoints, and object categories, so the reported benchmark results should not be assumed to generalize to all settings. Spatial scene representations could be misused for surveillance or deployed in safety-critical navigation without sufficient validation; such uses are outside the scope of this work.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training objectives, visual-cue trace construction, and inference procedure in Sec. 3 and Appendix A.1. The Experiments section and appendix identify the Qwen3-VL backbone (Bai et al., 2025a), the training and evaluation datasets, frame sampling, optimization settings, component configurations, and benchmark metrics. We also report separate MMI and VCLR ablations, component analyses, and geometric reconstruction results to support verification of their respective contributions. The underlying datasets and pretrained models are third-party resources and remain subject to their original access and licensing terms. We will publicly release the source code upon acceptance of the paper.

REFERENCES

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, pp. 1534–1543, 2016.

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, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025.

Meng Cao, Haokun Lin, Haoyuan Li, Haoran Tang, Rongtao Xu, Dong An, Xue Liu, Ian Reid, and Xiaodan Liang. Seeing through imagination: Learning scene geometry via implicit spatial world modeling. *arXiv preprint arXiv:2512.01821*, 2025.

Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. In *CVPR Findings*, pp. 7176–7187, June 2026.

Haijier Chen, Bo Xu, Shoujian Zhang, Haoze Liu, Jiaxuan Lin, and Jingrong Wang. Vid-llm: A compact video-based 3d multimodal llm with reconstruction-reasoning synergy. *arXiv preprint arXiv:2509.24385*, 2025a.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024.

Yiming Chen, Zekun Qi, Wenyao Zhang, Xin Jin, Li Zhang, and Peidong Liu. Reasoning in space via grounding in the world. *arXiv preprint arXiv:2510.13800*, 2025b.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025c.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *NeurIPS*, 35:5982–5994, 2022.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. *arXiv preprint arXiv:2501.01163*, 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Xiangjun Gao, Zhensong Zhang, Dave Zhenyu Chen, Songcen Xu, Long Quan, Eduardo Pérez-Pellitero, and Youngkyoon Jang. Map2thought: Explicit 3d spatial reasoning via metric cognitive maps. *arXiv preprint arXiv:2601.11442*, 2026.

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.

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. *NeurIPS*, pp. 20482–20494, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023a.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023b.

Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In *CVPR*, pp. 24570–24581. IEEE, 2025a.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. In *NeurIPS*, 2025b.

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.

Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning. *arXiv preprint arXiv:2410.00255*, 2024.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025a.

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

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. *arXiv preprint arXiv:2512.21218*, 2025b.

Dingkang Liang, Cheng Zhang, Xiaopeng Xu, Jianzhong Ju, Zhenbo Luo, and Xiang Bai. Cook and clean together: Teaching embodied agents for parallel task execution. *arXiv preprint arXiv:2511.19430*, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*, 2025.

Zhaochen Liu, Limeng Qiao, Guanglu Wan, and Tingting Jiang. Geoalign: Geometric feature realignment for mllm spatial reasoning. *arXiv preprint arXiv:2604.12630*, 2026.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Prisacariu. Do 3d large language models really understand 3d spatial relationships? In *ICLR*, volume 2026, pp. 32591–32610, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *ICLR*, 2023.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *arXiv preprint arXiv:2406.04334*, 2024.

Cheolhong Min, Jaeyun Jung, Daeun Lee, Hyeonseong Jeon, Yu Su, Jonathan Tremblay, Chan Hee Song, and Jaesik Park. Why far looks up: Probing spatial representation in vision-language models. *arXiv preprint arXiv:2605.30161*, 2026.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *ICRA*, pp. 6892–6903. IEEE, 2024.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *ICCV*, pp. 20076–20086. IEEE, 2023.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. In *ICLR*, volume 2026, pp. 157277–157302, 2026.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pp. 12159–12168. IEEE, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, pp. 10892–10902. IEEE, 2021.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, pp. 8216–8223. IEEE, 2023.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025.

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.

Chongyu Wang, Ting Huang, Chunyu Sun, Xinyu Ning, Di Wang, and Hao Tang. Let geometry guide: Layer-wise unrolling of geometric priors in multimodal llms. *arXiv preprint arXiv:2604.05695*, 2026a.

Haibo Wang et al. Learning geometric representations from videos for spatial intelligent multimodal large language models. *arXiv preprint arXiv:2606.05833*, 2026b.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. *arXiv preprint arXiv:2504.01901*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025b.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pp. 20697–20709. IEEE, 2024.

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

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. Spatiolm: Towards general physical spatial intelligence in vision-language models. *arXiv preprint arXiv:2608.01899*, 2026a.

Xianjin Wu, Dingkang Liang, Tianrui Feng, Kui Xia, Yumeng Zhang, Xiaofan Li, Xiao Tan, and Xiang Bai. Generation models know space: Unleashing implicit 3d priors for scene understanding. In *ECCV*, 2026b.

Beining Xu, Siting Zhu, Zhao Jin, Junxian Li, and Hesheng Wang. S2-mllm: Boosting spatial reasoning capability of mllms for 3d visual grounding with structural guidance. *arXiv preprint arXiv:2512.01223*, 2025.

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. *arXiv preprint arXiv:2412.14171*, 2025a.

Jihan Yang, Zifan Zhao, Xichen Pan, Shusheng Yang, Junyi Zhang, Bingyi Kang, Hu Xu, Shang-Wen Li, and Saining Xie. Cambrian-p: Pose-grounded video understanding. *arXiv preprint arXiv:2605.22819*, 2026.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025c.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.

Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, Guibin Zhang, Jiale Tao, Jiayi Zhang, et al. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*, 2026.

Jiang 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*, 2026a.

Jiaxin Zhang, Junjun Jiang, Haijie Li, Youyu Chen, Kui Jiang, and Dave Zhenyu Chen. Gap-mllm: Geometry-aligned pre-training for activating 3d spatial perception in multimodal large language models. *arXiv preprint arXiv:2603.16461*, 2026b.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024.

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

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space, 2025.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, pp. 2911–2921, 2023.

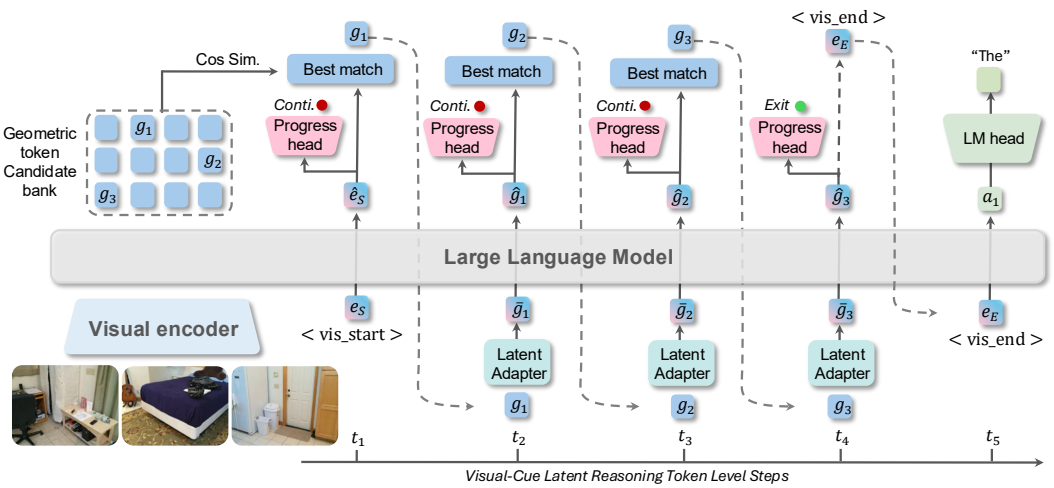


Figure 7: Inference with Visual-Cue Latent Reasoning. At each step, the model matches its predicted latent embedding to the geometric token bank and feeds the selected token into the next step. The progress head ends latent reasoning before the model generates a textual answer.

A Appendix

A.1 Inference Details

Figure 7 illustrates inference with Visual-Cue Latent Reasoning (VCLR). At step k, the LLM predicts a latent embedding $\widehat{g}_k$. For the geometry-aware hidden-state bank used in our method, let $\mathbf{G} = \{g_j\}_{j=1}^{N}$ denote the candidate tokens and $s_k(j) = \cos(\widehat{g}_k, g_j)$ their cosine similarity to the prediction. The evaluator can alternatively match against visual input embeddings and independently select input embeddings or final-layer visual hidden states for feedback. When configured, the latent adapter maps hidden-state feedback into the LLM input space, and a step embedding is applied to the feedback before the next reasoning step.

We design three decoding strategies (`greedy`, `no_repeat`, and `beam_search`), a separate feedback-source option `no_match` for retrieval-free latent decoding , and an optional spatial penalty that modifies retrieval scores. These options control the latent spatial reasoning trajectory before answer generation.

Greedy decoding (`greedy`). At each step, greedy decoding selects the token with the highest score s and feeds back its embedding. Previously selected tokens remain eligible and may therefore be retrieved more than once.

Greedy decoding without repetition (`no_repeat`). This strategy masks previously selected bank indices and greedily selects the highest-scoring eligible token. Each index can be selected at most once, and retrieval terminates when no eligible indices remain. We use this strategy by default.

Retrieval-free latent decoding (`no_match`). This strategy feeds the predicted state $\widehat{g}_i$ through the latent adapter to condition the next reasoning step directly.

Beam search. Beam search maintains up to B latent trajectories. For each active path at step k, let $\mathcal{C}_k$ denote its eligible bank indices and $s_k(j)$ the matching score with the j-th geometric token. Scores are normalized over all eligible indices to obtain transition log probabilities:

$$\log p_k(j) = \log \frac{\exp(s_k(j)/T)}{\sum_{i \in \mathcal{C}_k} \exp(s_k(i)/T)}, \qquad A = \frac{\sum_{k=1}^{L} \log p_k(j)}{\max(1, L)^{\alpha}}. \tag{10}$$

At each step, the decoder expands each active path with up to B candidates and retains the B extensions with the highest cumulative log probability, without length normalization during pruning. L denotes the length of VCLR path. Each path maintains its own retrieval history and stopping state. Completed paths are ranked by the length-adjusted score A, and the best path is replayed from the

prompt before answer generation. With $B = 1$, beam search reduces to greedy decoding under the same repetition and spatial-penalty settings.

Spatial penalty. An optional spatial penalty discourages repeated retrieval from a small region of one frame. Let x_j denote the center of the geometric token g_j on the two-dimensional patch grid after spatial merging, where adjacent patches are one unit apart, and let f_j denote its frame identifier. The penalty activates only when the last P selections are from the same frame and their centers are pairwise within distance r. This penalty is computed as:

$$\Delta_k(j) = \lambda\, \mathbb{I}[f_j = f_{j_{k-1}}] \max\left(0, 1 - \frac{\|x_j - x_{j_{k-1}}\|_2^2}{r^2}\right), \qquad \tilde{s}_k(j) = s_k(j) - \Delta_k(j). \tag{11}$$

The penalty is centered on the most recent selection and recomputed from a sliding window, so it does not accumulate across earlier visits. It applies only within the same frame and uses patch-grid distance.

VCLR Stopping and answer generation. All strategies share the configured stopping policy. With the progress head, the latent trajectory ends when the $\mathrm{P}_k(\varnothing)$ reaches the completion threshold τ. After stopping, the decoder consumes the cue-end token and generates the textual answer autoregressively.

A.2 Datasets and Benchmarks

VSI-Bench is a video-based spatial reasoning benchmark for multimodal large language models, constructed from the validation sets of three indoor 3D scene reconstruction datasets: ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021), contributing 88, 50, and 150 videos, respectively. It contains 288 real-world videos and over 5,000 question–answer pairs, using object-level 3D annotations to construct questions that models answer from video alone. The benchmark covers eight tasks across three categories: configurational tasks, including object counting, relative distance, relative direction, and route planning; measurement estimation tasks, including object size, room area, and absolute distance; and a spatiotemporal task involving first appearance ordering. Multiple-choice tasks are evaluated using accuracy (ACC), while numerical-answer tasks use Mean Relative Accuracy (MRA), which averages accuracy across relative-error thresholds from 5% to 50% in 5% increments to measure how closely predictions match ground-truth values.

VSI-590K is a large-scale instruction-tuning dataset introduced in Cambrian-S (Yang et al., 2025c) to enhance visual-spatial understanding in MLLMs. It comprises 590,667 question–answer pairs spanning 5,963 videos and 44,858 images from 10 sources: five 3D-annotated real-video datasets (S3DIS (Armeni et al., 2016), ADT (Pan et al., 2023), ScanNet, ScanNet++ V2, and ARKitScenes), two simulated datasets (ProcTHOR (Deitke et al., 2022) and Hypersim (Roberts et al., 2021)), and three sources of unannotated real videos (YouTube Room Tour, Open X-Embodiment (O'Neill et al., 2024), and AgiBot-World (Bu et al., 2025)), from which frames are extracted and pseudo-annotated. The dataset covers 12 question types across five categories: direction (43.1%), distance (30.4%), size (16.4%), counting (7.4%), and appearance order (2.7%). These tasks involve relative comparisons, absolute measurements, and spatial reasoning from camera and object perspectives. Question–answer pairs are automatically generated using diverse templates grounded in 3D annotations or reconstructed geometry, providing supervision for spatial configuration understanding, metric estimation, and temporal reasoning.

VLM-3R-207K (Fan et al., 2025) is a spatial instruction-tuning dataset that contains 207,779 question–answer pairs derived from ScanNet, ScanNet++, and ARKitScenes. It covers seven tasks: relative direction (86,441), absolute distance (50,757), relative distance (42,025), object size estimation (12,917), object counting (9,357), route planning (4,225), and room size estimation (2,057). The questions are automatically generated from structured scene metadata containing 3D object geometry, semantic annotations, and camera poses, while route-planning examples are constructed using the Habitat simulator. This dataset provides supervision for learning spatial relationships, metric estimation, and embodied navigation from monocular videos.

MindCube (Yin et al., 2025) evaluates spatial reasoning from limited multi-view observations, comprising 21,154 question–answer pairs and 3,268 images organized into 976 multi-view groups. It covers three camera movement settings: Rotation (1,081 questions), Among (18,204 questions), and Around (1,869 questions), requiring models to integrate information across viewpoints, perform

perspective taking, and reason about occluded or out-of-view objects. MindCube-Tiny is a diagnostic evaluation subset containing 1,050 questions, including 200 Rotation, 600 Among, and 250 Around questions. **MindCube-Train** provides 10,000 question–answer pairs with corresponding cognitive maps and free-form reasoning chains for supervised fine-tuning, supporting the learning of explicit spatial representations and reasoning processes.

Real-3DQA (Ma et al., 2026) is a diagnostic benchmark derived from SQA3D (Ma et al., 2023) to evaluate 3D spatial reasoning while reducing the influence of linguistic shortcuts. It retains 1,485 of the original 3,519 test questions through filtering based on comparisons between 3D-LLMs and their text-only counterparts, supplemented by GPT-based filtering. The benchmark covers 11 question categories, including spatial relations, measurement, counting, object identification, visibility, and navigation. It additionally provides a viewpoint-rotation evaluation set comprising 750 question groups and 3,000 question–answer pairs, with each group covering the original viewpoint and rotations of 90°, 180°, and 270°. These augmented examples undergo manual validation and expert review. Evaluation combines exact-match metrics with a Viewpoint-Rotation Score (VRS) to assess answer accuracy across changes in the observer's orientation.

A.3 Training Settings

To ensure a fair comparison, we adopt the training settings of the corresponding baseline methods on VSI-bench, MindCube, and Real-3DQA. For VSI-bench, we follow GeoVR (Wang et al., 2026b) and train on VSI-590K and VLM-3R for one epoch. For MindCube, we follow SpatialSV (Tang et al., 2026) and fine-tune the model using 10K training samples from MindCube (Yin et al., 2025). For Real-3DQA, we follow its evaluation protocol (Ma et al., 2026) and fine-tune the model on the SQA3D (Ma et al., 2023) training set for one epoch.

Tables 5–7 summarize the training configurations, including data preparation, optimization, and the parameters of Mental Map Imaging (MMI) and Visual-Cue Latent Reasoning (VCLR). We jointly train the LLM, multimodal projector, and task-specific modules while keeping the visual encoder frozen.

Table 5: General training settings for Artemis-3D on VSI-Bench.

Setting	**Value**
Model and data	
Base MLLM	Qwen3-VL-2B-Instruct
Training dataset	VSI-590K, VLM-3R
Frames per scene	$4 \sim 32$
Trainable backbone components	LLM and multimodal projector
Visual encoder	Frozen
Optimization	
Training epochs	1
Global / per-device batch size	32 / 1
LLM learning rate	2×10^{-5}
Multimodal projector learning rate	1×10^{-5}
Geometry / metric-scale learning rate	1×10^{-4} / 1×10^{-4}
VCLR / progress-head learning rate	1×10^{-4} / 1×10^{-4}
Learning-rate schedule	Cosine decay
Warmup ratio	0.03
Weight decay	0

Table 6: Mental Map Imaging configuration. Decoder-layer indices are zero-based. Loss weights are listed separately for each reconstruction objective.

Setting	Value
Depth head	
Intermediate layers	5, 13, 20, 27
Regression loss	ℓ_1
Depth loss weight	1.0
Regression / gradient loss weights	1.0 / 1.0
Gradient scales	1, 2, 4, 8
Confidence filtering ratio	0.0
Point-map head	
Intermediate layers	5, 13, 20, 27
Regression loss	Smooth ℓ_1
Point-map loss weight	1.0
Confidence filtering ratio	0.2
Confidence regularization coefficient	0.2
Camera head	
Head dimension / depth / attention heads	256 / 1 / 16
Refinement iterations	4
Camera loss weight	2.5
Translation / rotation / field-of-view weights	1.0 / 1.0 / 0.5
Iterative-stage weighting factor	0.6
Minimum trajectory length for balancing	0.1
Others	
Scene tokens	32
Metric-scale loss weight	1.0

Table 7: Visual-Cue Latent Reasoning configuration.

Setting	Value
Evidence selection and supervision	
Maximum regions per evidence group	4
Token selection	Box center, radius 1 (up to 3×3 tokens)
Maximum visual-cue length	256
Token prediction loss / weight	Cosine distance / 1.0
Progress and stopping	
Maximum remaining-group count	8
Progress loss weight	0.1
Completion-class weight	2.0
Count / monotonicity loss weights	0.05 / 0.01
Stop loss weight	0.1
Continue-loss balancing weight	0.5
Initial stopping probability	0.01
Completion threshold τ	0.8

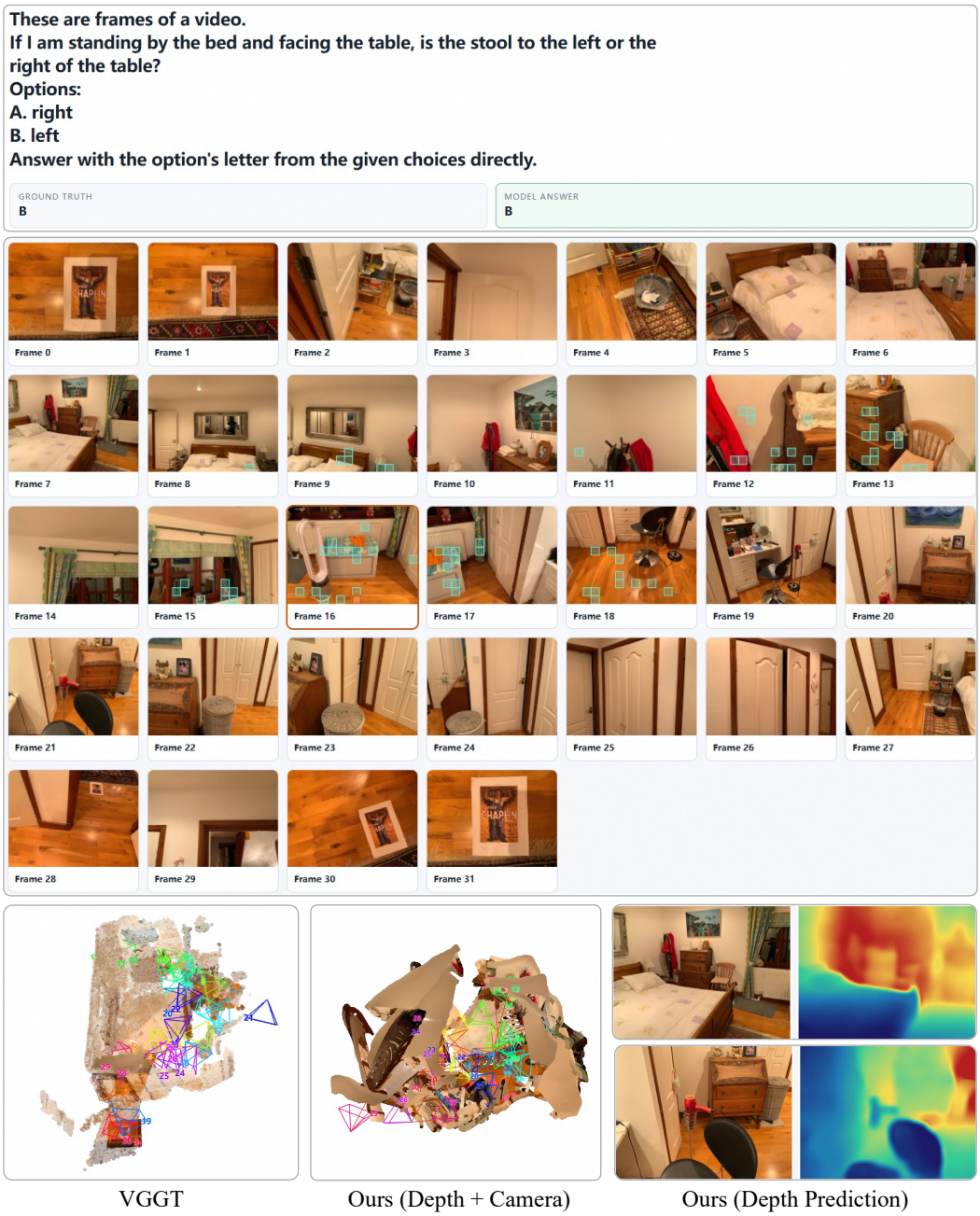


Figure 8: Viewpoint-conditioned spatial reasoning example. When asked to stand by the bed and face the table, Artemis-3D correctly identifies the stool as being to the left of the table (option B). The visualization shows the selected visual cues across frames, scene reconstructions with camera poses, and representative frames with depth maps.

A.4 Qualitative Examples

Figures 9 and 8 show qualitative examples of object counting and viewpoint-conditioned spatial reasoning, including the input frames, selected visual cues, and geometric visualizations.

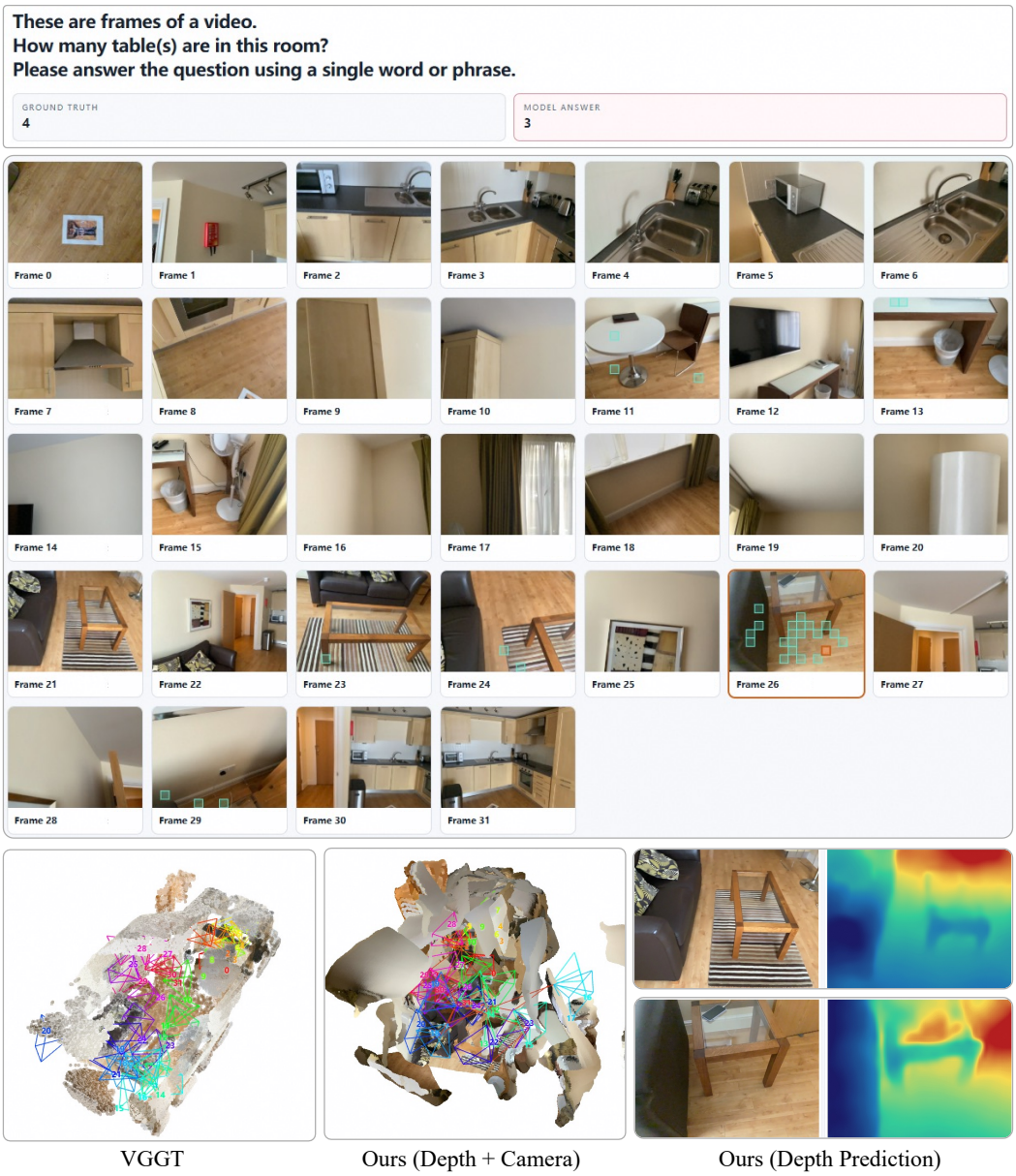


Figure 9: Object-counting example. Given a video of a room, Artemis-3D predicts three tables, whereas the ground-truth count is four. The visualization shows the selected visual cues across frames, scene reconstructions with camera poses, and representative frames with depth maps.

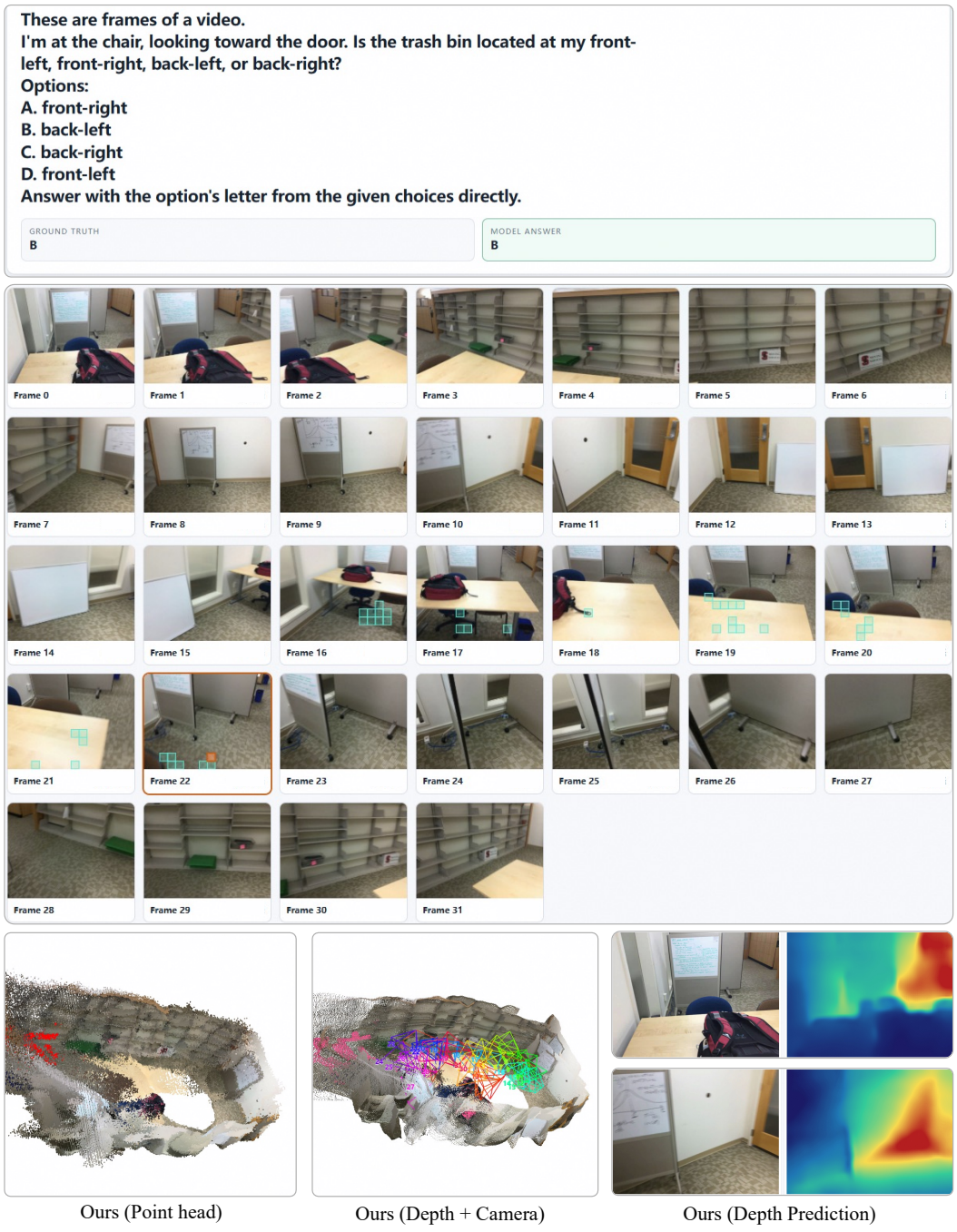


Figure 10: Object-counting example. Given a video of a room, Artemis-3D predicts three tables, whereas the ground-truth count is four. The visualization shows the selected visual cues across frames, scene reconstructions with camera poses, and representative frames with depth maps.