Latent Spatial Reasoning: Building Innate 3D Awareness via Latent-Space Distillation

Ruifei Zhang[1,2,3], **Xiangru Lin**[5], **Haoyuan Li**[6], **Wei Zhang**[4], **Xiao Tan**[4], **Junlin Xie**[1,3], **Xiang Wan**[3], **Guanbin Li**[2,7,†]
[1] The Chinese University of Hong Kong, Shenzhen
[2] Sun Yat-sen University [3] Shenzhen Research Institute of Big Data [4] Baidu Inc.
[5] The University of Hong Kong [6] Wuhan University [7]Shenzhen Loop Area Institute
ruifeizhang@link.cuhk.edu.cn liguanbin@mail.sysu.edu.cn

Abstract

Humans effortlessly infer 3D structure, such as depth, occlusion, and spatial arrangement, from 2D images and reason about it fluidly. Multimodal large language models still struggle with such spatial reasoning. Current approaches attempt to bridge this gap by injecting explicit 3D priors at inference time or aligning features with 3D-aware teachers during training. While these strategies improve geometric perception, they typically treat 3D knowledge as static representations, rather than enabling the model to reason with 3D cues during inference. We argue that closing this gap requires internalizing 3D knowledge in a semantically aligned form that can participate in the model's reasoning dynamics. To this end, we introduce **Latent Spatial Reasoning (LSR)**, a framework that distills geometric knowledge into *latent spatial tokens*: geometry-aware representations that the model actively queries and conditions on to guide its spatial reasoning process. Through hierarchical distillation, LSR transfers 3D knowledge from foundation models, aligns it with the vision-language space, and trains the model to interleave spatial and linguistic reasoning, all without external 3D modules at inference time. Experiments on diverse spatial reasoning benchmarks demonstrate state-of-the-art performance.

1 Introduction

Multimodal large language models (MLLMs) have achieved remarkable success in 2D vision-language understanding [60, 27, 2], yet they remain far from human-level capability in 3D spatial reasoning: inferring depth, occlusion, and object arrangement from 2D images [45]. Human spatial cognition involves not only reconstructing 3D structure from visual input but also leveraging that mental representation to draw inferences. This distinction suggests two competencies that 3D-aware MLLMs should possess: (1) **constructing** internal 3D representations from 2D inputs, and (2) **reasoning with** those representations dynamically during inference. As illustrated in Figure 1, most existing methods either fall short in both capabilities or address primarily the former while leaving the latter largely unchanged. *Architectural augmentation* methods [17, 59, 56] incorporate explicit 3D priors (depth maps, point clouds, or 3D foundation model outputs) at inference time, improving geometric perception but imposing computational overhead and relying on external modules decoupled from the model's reasoning dynamics. *Representation supervision* methods [18, 23] align 2D features with 3D-aware foundation models during training, enriching representations but providing no mechanism to actively engage them during inference. In both paradigms, 3D knowledge remains static, serving as an auxiliary input or a frozen embedding, rather than an integrated component that actively participates in the model's reasoning process.

† Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

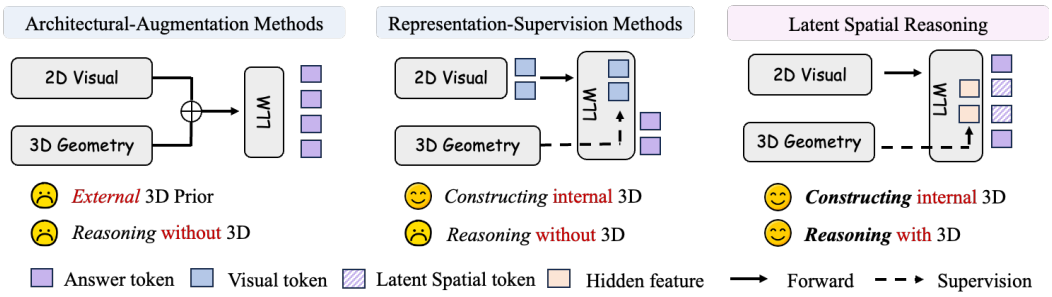


Figure 1: The illustration of our latent spatial reasoning (LSR). Through hierarchical distillation, LSR internalizes 3D knowledge into geometry-aware and semantically aligned latent spatial tokens, enabling the model to effectively construct and leverage such 3D priors during spatial reasoning.

The emergence of latent reasoning frameworks offers a promising direction for addressing this bottleneck [32, 20, 49]. Recent work has shown that models can perform complex inference through a small set of latent variables that implicitly encode intermediate reasoning states, suggesting the possibility of embedding 3D spatial understanding directly within a model's cognitive process rather than treating it as an external input. Yet extending latent reasoning to 3D space is far from straightforward. While aligning latent tokens with the feature space of a 3D foundation model can endow them with geometric information [8], the representational gap between geometric structure and language semantics makes it difficult to fully incorporate 3D cues into reasoning chains during inference, limiting the effective guidance of 3D information for subsequent textual reasoning. These difficulties underscore the need for a framework that unifies 3D representation construction and 3D-aware reasoning within a single semantically aligned latent workspace.

We posit that bridging the geometric–semantic gap requires not only injecting 3D knowledge into the model, but structurally anchoring it within the model's existing representational space. Simply infusing 3D-aware features risks decoupling them from the vision–language semantics that underlie reasoning, leading to performance degradation (see Table 5). Instead, 3D knowledge must be both geometry-aware and semantically fused, encoded in a form that preserves spatial discriminability while remaining native to the model's linguistic dynamics. Only then can it guide reasoning seamlessly, without disrupting the coherence of multimodal understanding.

We materialize this principle in **Latent Spatial Reasoning (LSR)**, a framework that transforms 3D understanding from a static representational property into a dynamic reasoning capability. At its core, LSR establishes an *internal spatial workspace* through **latent spatial tokens**: learnable, geometry-aware and semantically-aligned representations that encode 3D structural cues. Crucially, this workspace operates entirely within the model's latent dynamics, requiring no external 3D modules at inference time. To construct this capability, we design a hierarchical **latent-space distillation** procedure comprising complementary components: (i) *3D Concept Distillation* transfers fundamental geometric knowledge from pretrained 3D foundation models into the latent spatial tokens; (ii) *Intermediate Feature Alignment* bridges the geometric–semantic gap by aligning latent tokens with a geometrically-aligned teacher model trained for robust 2D–3D feature fusion, thereby ensuring that the distilled 3D cues remain fully compatible with the MLLM's vision–language space; and (iii) *On-Policy Reinforcement Learning* trains the model to actively interleave spatial and linguistic reasoning. Together, these stages enable LSR to construct, maintain, and reason over internal 3D representations, all within the MLLM's latent dynamics. Across diverse 3D spatial reasoning benchmarks, LSR achieves state-of-the-art performance, demonstrating the effectiveness of embedding spatial reasoning directly into the model's cognitive process.

In summary, our contributions are threefold:

- We identify a fundamental limitation of current spatial reasoning methods and articulate two complementary competencies: *constructing* semantically aligned 3D representations and *reasoning with* them dynamically, as requirements for 3D-aware MLLMs.
- We introduce Latent Spatial Reasoning (LSR), a framework that integrates 3D understanding directly into the reasoning process via latent spatial tokens and hierarchical latent-space distillation.
- We demonstrate that LSR achieves state-of-the-art performance across diverse 3D spatial reasoning benchmarks while operating purely from 2D inputs at inference time.

2 Related Work

2.1 3D Spatial Understanding

The pursuit of 3D spatial understanding in MLLMs has evolved along two interconnected tracks. The first focuses on architectural enhancements to ground MLLMs in 3D data: early works directly process point clouds [7, 17], integrate depth or 3D positional encodings into 2D vision features [59, 57], or leverage compressed Bird's-Eye-View representations [29, 36]. While these methods provide effective low-level 3D perceptual grounding, they primarily address explicit 3D pattern recognition rather than reasoning. The second track, propelled by challenging benchmarks such as VSI-Bench [45] and SPAR-Bench [52], aims to unlock higher-order spatial intelligence [6, 46, 47]. Existing works have diverged into two strategies: methods like VG-LLM [56] and Spatial-MLLM [41] leverage powerful 3D foundation models to provide rich scene priors, effectively outsourcing 3D understanding to external modules; alternatively, VST [46] and Cambrian-S [47] seek to reinforce endogenous spatial reasoning through tailored training data or architectural inductive biases. More recently, 3DThinker [8] explores latent 3D reasoning by internalizing geometric knowledge in latent tokens. However, how to make such geometric knowledge fully compatible with semantic reasoning remains insufficiently addressed. LSR addresses this challenge through hierarchical distillation, bridging the geometric-semantic gap and enabling geometry-aware, semantically aligned 3D reasoning.

2.2 Visual Reasoning

Chain-of-thought prompting has proven effective for eliciting complex reasoning in large language models [15, 44, 50]. A natural extension is to reason not only *about* images but *with* images, generating or manipulating visual content as intermediate reasoning steps via a series of tool functions. Methods such as DeepEyes [58], Pixel Reasoner [37], and Thyme [54] explore this direction, demonstrating that visual intermediate representations can enhance vision-language joint reasoning. However, these approaches are typically constrained to a predefined set of tool types, which limits their flexibility and scalability. A parallel line of work investigates *latent reasoning*, where models perform inference by a small set of continuous representations rather than a large sequence of discrete tokens. In the language domain, COCONUT [16] introduces continuous thought states derived from hidden layers, enabling breadth-first search over reasoning paths. CoDi [32] demonstrates that explicit chain-of-thought can be distilled into continuous representations via self-distillation, achieving comparable performance with significant compression. Recent works have extended latent reasoning to vision-language models [49, 20, 39]: Latent Visual Reasoning [20] enables autoregressive reasoning directly in the visual embedding space, while Monet [39] introduces a training framework for generating continuous embeddings as intermediate visual thoughts, addressing computational costs through three-stage distillation and Visual-latent Policy Optimization. Our approach builds on this foundation for 3D spatial reasoning: LSR introduces latent spatial tokens that encode geometric structure and are trained through hierarchical distillation to bridge the geometric–semantic gap and actively participate in inference.

2.3 Knowledge Distillation

Knowledge distillation has been widely adopted to transfer capabilities across models or compress reasoning processes into efficient forms. In the MLLM domain, prior work has explored distillation for model compression [13] and vision-language alignment. More relevant to our work is distillation in the context of latent reasoning: CoDi [32] demonstrates that explicit chain-of-thought can be compressed into continuous representations through self-distillation, enabling efficient reasoning without token-level verbalization. Monet [39] extends this to visual reasoning through a three-stage supervised fine-tuning pipeline that aligns latent embeddings with visual observations. LSR draws on these principles but addresses a distinct challenge: transferring structured 3D knowledge from geometric foundation models into latent representations that remain compatible with the MLLM's vision-language space. Our hierarchical distillation procedure, progressively builds the model's capacity to internalize and actively use geometric knowledge during reasoning.

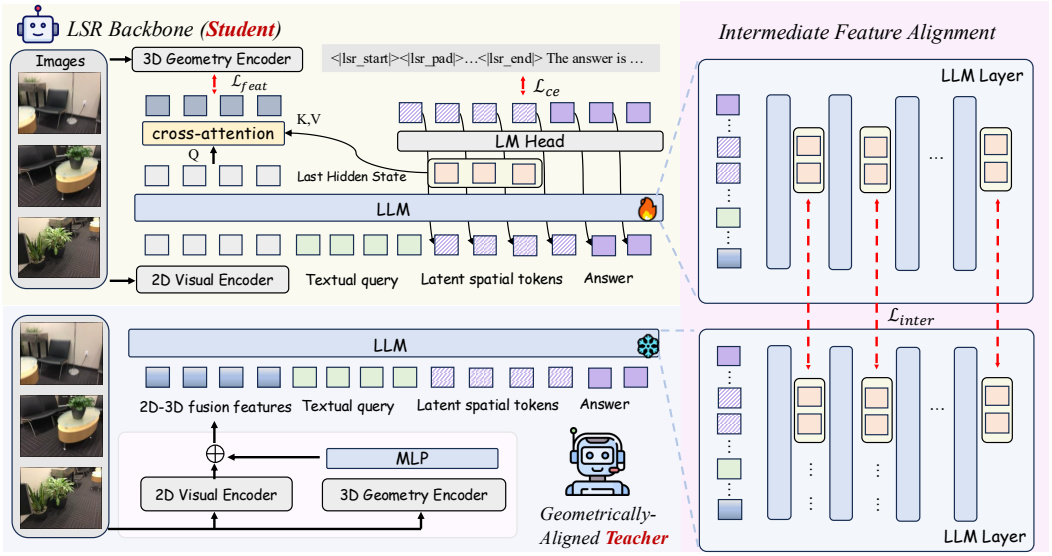


Figure 2: The overview of our LSR framework with hierarchical latent-space distillation. The framework introduces latent spatial tokens into MLLMs, enabling them to encode compact 3D concepts distilled from a 3D foundation model. A geometrically-aligned teacher further aligns the intermediate features of these tokens, ensuring seamless integration with the vision–language space and effective guidance for spatial reasoning.

3 Method

3.1 Overview

Our objective is to enable MLLMs to develop two complementary capabilities: *constructing* internal 3D representations from 2D inputs, and *reasoning* with those representations during inference without external 3D modules. To achieve this, we introduce the **Latent Spatial Reasoning (LSR)** framework, which transforms 3D understanding from a static representational property into a dynamic reasoning capability through hierarchical latent-space distillation, as illustrated in Figure 2. The framework comprises three complementary components. **3D Concept Distillation** (Sec. 3.3) serves as the LSR backbone, addressing the first capability by transferring geometric knowledge from pretrained 3D foundation models into compact *latent spatial tokens*, learnable representations that encode 3D structure within the model's hidden states. **Intermediate Feature Alignment** (Sec. 3.4) bridges the geometric–semantic gap by aligning these tokens with a teacher model trained for robust 2D–3D fusion, ensuring compatibility with the MLLM's vision–language dynamics. **On-Policy Reinforcement Learning** (Sec. 3.5) enhances the second capability, training the model to actively interleave spatial and linguistic reasoning so that 3D knowledge shapes the inference process.

3.2 Preliminaries

MLLM. A standard MLLM comprises a visual encoder $\mathcal{E}_{\text{2d}}$, a connector $\mathcal{P}_{\text{2d}}$, and a language model $\mathcal{M}$. Given input frames $\mathbf{I} = \{I_1, \ldots, I_N\}$ and a textual query $\mathbf{Q}$, the visual encoder extracts features projected into the language model's embedding space:

$$\mathbf{F}_{\text{2d}} = \mathcal{P}_{\text{2d}}(\mathcal{E}_{\text{2d}}(\mathbf{I})) \in \mathbb{R}^{NL_v \times d_h}, \tag{1}$$

where $L_v = \lfloor h/p \rfloor \times \lfloor w/p \rfloor$ denotes the number of visual tokens per frame, p is the patch size, and d_h is the hidden dimension. The LLM processes the concatenated visual and text tokens autoregressively to generate response $\mathbf{A} = \{a_1, \ldots, a_T\}$.

3D Foundation Model. We employ VGGT [38], a state-of-the-art 3D scene understanding model, as our geometric knowledge source. VGGT comprises an encoder $\mathcal{E}_{\text{3d}}$ and a connector $\mathcal{P}_{\text{3d}}$ that produce dense 3D-aware features:

$$\mathbf{F}_{\text{3d}} = \mathcal{P}_{\text{3d}}(\mathcal{E}_{\text{3d}}(\mathbf{I})) \in \mathbb{R}^{NL_v \times d_h}. \tag{2}$$

3.3 3D Concept Distillation

This stage addresses the *construction* capability: equipping the MLLM with the ability to build internal 3D representations. Inspired by recent advances in latent reasoning, a straightforward approach would insert special tokens whose hidden states are trained to regress toward 3D features extracted from a foundation model (e.g., VGGT), thereby injecting 3D knowledge into the MLLM. However, this confronts the **Supervision Efficiency** challenge: A 3D foundation model like VGGT produces dense features for each image. Directly supervising equally dense predictions imposes prohibitive overhead, while downsampling sacrifices fine-grained geometric detail. In our work, we introduce *3D concept distillation* as the backbone of LSR to address this dilemma.

Latent Spatial Tokens. Rather than directly regressing dense features, we design latent spatial tokens to serve as *conditional keys and values* that, through cross-attention with visual tokens, adaptively transform visual features into structured 3D-aware representations. Specifically, we extend the MLLM vocabulary with special tokens: a start token `[lsr_start]`, M placeholder tokens `[lsr_pad]`, and an end token `[lsr_end]`. The M placeholders function as latent spatial tokens encoding 3D concepts. Let $\mathbf{H}_{\text{3d}} \in \mathbb{R}^{M \times d_h}$ and $\mathbf{H}_{\text{2d}} \in \mathbb{R}^{NL_v \times d_h}$ denote the hidden states of these latent tokens and visual token features, respectively, where both are obtained from the final layer of the LLM. Dense 3D-aware features are reconstructed via cross-attention:

$$\hat{\mathbf{F}}_{\text{3d}} = \text{CrossAttn}(\mathbf{H}_{\text{2d}}, \mathbf{H}_{\text{3d}}, \mathbf{H}_{\text{3d}}) \in \mathbb{R}^{NL_v \times d_h}, \tag{3}$$

supervised by cosine similarity against VGGT features:

$$\mathcal{L}_{\text{feat}} = 1 - \text{cosine}(\hat{\mathbf{F}}_{\text{3d}}, \mathbf{F}_{\text{3d}}). \tag{4}$$

Semantic Grounding. Unlike prior works that directly feed hidden features into the reasoning chain, we anchor spatial tokens to the linguistic space by supervising each spatial token via cross-entropy loss $\mathcal{L}_{\text{ce}}$ to predict its vocabulary entry. This ensures tokens carrying 3D information remain valid linguistic units, integrating seamlessly into autoregressive generation. This grounding also establishes a foundation for the next *Intermediate Feature Alignment* stage, where vocabulary embeddings serve as the basis for reconstructing well-aligned and discriminative 3D knowledge.

3.4 Intermediate Feature Alignment

The preceding stage distills geometric knowledge into latent tokens, but a gap remains: these representations originate from VGGT's feature space, not the MLLM's vision-language space. A model might accurately encode depth relationships yet fail to leverage this knowledge when answering spatial questions—geometric and linguistic reasoning operate in different representational subspaces. The core objective of this stage is to align the intermediate hidden features of the latent tokens with those of a geometrically-grounded teacher by enforcing representation consistency across LLM layers.

Geometrically-Aligned Teacher. We first train a teacher model designed for robust 3D-grounded vision-language understanding via explicit 2D–3D feature fusion. In contrast to our LSR backbone, which learns to predict 3D concepts intrinsically, the teacher model incorporates external 3D structural features extracted from the VGGT encoder. Formally, the teacher extracts 2D visual tokens $\mathbf{F}_{\text{2d}}$ and 3D geometric tokens $\mathbf{F}_{\text{3d}}$ in parallel, fusing them via an additional learned projector:

$$\mathbf{F}_{\text{tea}} = \mathbf{F}_{\text{2d}} + \text{MLP}(\mathbf{F}_{\text{3d}}). \tag{5}$$

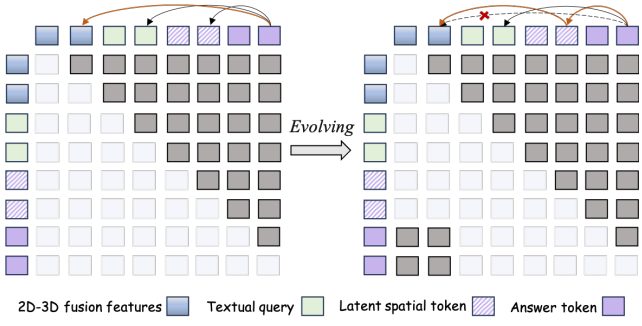


Figure 3: The illustration of progressive attention modulation.

This fused representation explicitly combines appearance and geometry, providing 3D-grounded input to the teacher's LLM.

Progressive Attention Modulation. While the teacher model achieves effective 2D-3D feature fusion, a critical question remains: *how can we leverage the teacher's aligned representations to guide the student's latent spatial tokens?* Our strategy is to equip the teacher with the same M latent tokens, denoted as $\mathbf{S}_{\text{tea}}$, forming training sequence $[\mathbf{F}_{\text{tea}}, \mathbf{Q}, \mathbf{S}_{\text{tea}}, \mathbf{A}]$, with latent tokens positioned between visual features and answer tokens. The core challenge then becomes ensuring that these tokens effectively condense the aligned multimodal representations into a geometry-rich knowledge carrier that is readily transferable during distillation. We achieve this through *progressive attention modulation*, as illustrated in Figure 3: gradually reducing attention from answer tokens to visual features forces answer generation to increasingly rely on the latent tokens. Let $\alpha(\tau) \in [0, 1]$ decay linearly over training:

$$\alpha(\tau) = \max\left(0, 1 - \frac{\tau}{\tau_{\text{total}}}\right), \tag{6}$$

where τ, τ_{total} are the current and total training steps, respectively. The modulated attention for answer token i attending to visual token j is:

$$\tilde{\text{Attn}}_{ij}^{(l)} = \begin{cases} \alpha(\tau) \cdot \text{Attn}_{ij}^{(l)} & \text{if } i \in \mathbf{A},\ j \in \mathbf{F}_{\text{tea}} \\ \text{Attn}_{ij}^{(l)} & \text{otherwise} \end{cases} \tag{7}$$

Crucially, latent tokens maintain full attention to visual features throughout, enabling them to absorb aligned 2D–3D knowledge. Through this modulation, the teacher's latent tokens evolve into *information bottlenecks* that comprehensively encode geometric-semantic representations which are ideally suited for subsequent distillation to the student model.

Distillation. We transfer knowledge by aligning intermediate representations across K selected LLM layers. Let $\mathbf{H}_{\text{tea}}^{(l)}$ and $\mathbf{H}_{\text{stu}}^{(l)}$ denote hidden states of latent tokens at layer l, the total distillation loss is the average over all selected layers:

$$\mathcal{L}_{\text{inter}} = \frac{1}{K} \sum_{k=1}^{K} \left(1 - \text{cosine}(\mathbf{H}_{\text{stu}}^{(l)}, \mathbf{H}_{\text{tea}}^{(l)})\right). \tag{8}$$

Thus, the complete SFT objective is:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{feat}} + \beta \mathcal{L}_{\text{inter}}, \tag{9}$$

where λ and β control distillation strength.

3.5 On-Policy Reinforcement Learning

The preceding stages build aligned 3D representations, addressing the *construction* capability. This stage enhances the *reasoning* capability: learning to actively engage its spatial awareness and reason with 3D cues during inference. To this end, we adopt a fully on-policy Group Relative Policy Optimization (GRPO) [31, 43], which refines the reasoning strategy through interactive, reward-driven policy updates.

On-Policy Trajectory Generation. At each training step, our LSR backbone interacts with a spatial reasoning task by auto-regressively generating a group of G trajectories $\{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_G\}$ from its policy $\pi_\theta(\cdot \mid \mathbf{I}, \mathbf{Q})$. Each trajectory $\boldsymbol{\tau}_i = (\hat{a}_1, \ldots, \hat{a}_T)$ is a complete token sequence that interleaves latent spatial tokens with ordinary text tokens, reflecting the spatial–linguistic reasoning process.

Policy Optimization. The policy is updated by maximizing the following GRPO objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\boldsymbol{\tau} \sim \pi_\theta}\left[\frac{1}{\sum_{i=1}^{G} |\boldsymbol{\tau}_i|} \sum_{i=1}^{G} \sum_{t=1}^{|\boldsymbol{\tau}_i|} A_{i,t}\right], \tag{10}$$

where $A_{i,t}$ denotes the advantage computed from the group-relative reward:

$$A_{i,t} = \frac{R(\boldsymbol{\tau}_i) - \mu_R}{\sigma_R}. \tag{11}$$

Table 1: Comparison results on video-based VSI-Bench.

Methods	Rank	Avg.	Numerical Question				Multiple-Choice Question			
			Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Proprietary Models (API)										
GPT-4o [19]	3	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash [34]	2	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro [34]	1	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-sourced VLMs										
InternVL2-8B [9]	6	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B [9]	4	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
Qwen2.5VL-7B [2]	7	33.0	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8
Qwen2.5VL-72B [2]	3	37.0	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9
LLaVA-OneVision-7B [21]	8	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B [21]	2	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-NeXT-Video-7B [26]	5	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B [26]	1	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Spatial-Enhanced Models										
SpatialLadder-3B [24]	7	45.7	63.5	34.3	61.7	43.9	45.4	44.8	35.6	36.4
Spatial-MLLM-4B [41]	5	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VG-LLM-4B [56]	6	47.3	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4
VST-3B [46]	3	57.7	66.6	45.0	72.8	60.9	59.9	47.6	40.7	68.3
3DThinker-3B [8]	2	59.1	-	-	-	-	-	-	-	-
Cambrian-S-3B [47]	4	57.3	70.7	40.6	68.0	46.3	64.8	61.9	27.3	78.8
LSR-3B (Ours)	1	61.5	68.5	44.6	71.7	65.4	58.4	65.1	41.2	76.8
VG-LLM-8B [56]	5	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
SpaceR-7B [28]	6	45.6	-	-	-	-	-	-	-	-
3DThinker-7B [8]	2	63.7	-	-	-	-	-	-	-	-
VLM-3R-7B [12]	4	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
VST-7B [46]	3	61.2	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2
LSR-7B (Ours)	1	64.5	70.0	46.5	70.9	65.1	62.5	82.0	42.8	75.9

Here, $R(\boldsymbol{\tau}_i)$ evaluates both answer correctness and format adherence, ensuring that the response is both factually accurate and follows the predefined output structure. The terms μ_R and σ_R denote the mean and standard deviation of the rewards across the G sampled trajectories within the same group. At this stage, we intentionally avoid imposing explicit constraints on the latent spatial token features, preserving the model's autonomy to dynamically adjust and integrate these representations in service of joint spatial–linguistic reasoning.

4 Experiments

We conduct comprehensive experiments to address the following research questions: **RQ1:** Does LSR achieve state-of-the-art performance on spatial reasoning benchmarks? **RQ2:** How does each distillation stage contribute to the overall performance? **RQ3:** What is the most effective strategy for distilling dense 3D features into compact latent tokens? **RQ4:** How should 3D priors be integrated into the reasoning process? **RQ5:** How do teacher-student architectural choices affect distillation quality? **RQ6:** How do hyperparameters (latent token size, alignment layers) influence performance?

4.1 Experimental Setup

Datasets and Benchmarks. Our training set is curated from multiple sources, including SPAR-7M [52], VSI-590K [47], and the LLaVA-Hound split [55], resulting in approximately 400K samples. We evaluate on benchmarks spanning three categories: single-image reasoning (CVBench [35]), multi-view image reasoning (SPAR-Bench [52], MMSI-Bench [48], ViewSpatial [22]), and video-based reasoning (VSI-Bench [45]).

Implementation Details. Our framework is built upon Qwen2.5-VL (3B and 7B variants) with VGGT as the 3D foundation model. We set the latent spatial token size $M = 4$ and configure K as half of the total LLM layers, applying uniformly-strided alignment for intermediate feature distillation with a fixed stride of 2. Further details are provided in the Appendix.

4.2 RQ1: State-of-the-Art Comparison

Video-based Spatiotemporal Reasoning. Table 1 presents results on VSI-Bench. Among proprietary APIs, Gemini-1.5 Pro achieves 45.4%, while the strongest open-source baseline (LLaVA-NeXT-Video-72B) reaches 40.9%. Spatially-enhanced models show substantially stronger performance through various strategies: architectural augmentation with external 3D encoders (Spatial-MLLM, VG-LLM, VLM-3R), specialized training data curation, or dedicated reasoning modules. Our LSR-

Table 2: Comparison results on single-image and multi-view spatial reasoning benchmarks.

Methods	Rank	Avg.	CVBench			SPAR-Bench	MMSI-Bench	ViewSpatial
			2D	3D	Avg.			
Spatial-Enhanced Models								
SpatialLadder-3B [24]	5	43.8	72.4	74.9	73.7	34.4	27.4	39.8
Spatial-MLLM-4B [41]	7	42.4	-	-	73.8	35.3	26.1	34.6
SpaceR-7B [28]	4	44.0	-	-	75.3	37.6	27.4	35.8
ViLaSR-7B [42]	3	45.1	-	-	77.1	37.6	30.2	35.7
Cambrian-S-3B [47]	6	43.1	-	-	75.2	33.0	25.2	39.0
Cambrian-S-7B [47]	2	45.3	-	-	76.9	37.9	25.8	40.9
LSR-7B (Ours)	1	51.4	69.8	88.3	79.0	50.8	31.5	44.3

Table 3: Ablation of each distillation component.

ID	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{inter}}$	$\mathcal{L}_{\text{grpo}}$	Numerical	Multiple-Choice	**Avg.**
1	✗	✗	✗	59.5	52.6	56.1
2	✓	✗	✗	60.5	58.1	59.4
3	✓	✓	✗	61.5	59.6	60.6
4	✓	✓	✓	62.5	60.3	61.5

Table 4: Ablation on 3D concept distillation strategies.

ID	Method	Numerical	Multiple-Choice	**Avg.**
1	Compression	59.7	57.1	58.5
2	Denoising	60.6	57.6	59.1
3	Cross-Attention	60.5	58.1	59.4

3B and LSR-7B achieve 61.5% and 64.5% respectively, establishing new state-of-the-art performance. Notably, LSR accomplishes this without external 3D modules at inference time, demonstrating that latent-space distillation effectively transfers 3D reasoning capability into the model's internal dynamics.

Multi-View Scene Understanding. Table 2 summarizes results on SPAR-Bench, MMSI-Bench, and ViewSpatial. LSR-7B attains leading performance across all three benchmarks, demonstrating consistent effectiveness in spatial reasoning from multi-view inputs. The gains are particularly pronounced on SPAR-Bench (+12.9% over the next best method), suggesting that our latent spatial tokens effectively capture cross-view geometric relationships.

Single-Image Spatial Reasoning. We further evaluate on CVBench to assess whether 3D-focused training compromises 2D perception. As shown in Table 2, LSR-7B achieves the best average performance while maintaining competitive 2D scores (69.8%), indicating that our hierarchical distillation preserves general visual understanding while enhancing spatial reasoning.

4.3 RQ2: Contribution of Each Distillation Stage

Table 3 isolates the contribution of each stage. Starting from Qwen2.5-VL-3B with standard cross-entropy training (ID #1, a plain SFT baseline), we progressively add components: **+3D Concept Distillation (ID #2):** Introduces geometric awareness by distilling VGGT features into latent tokens, improving performance from 56.1% to 59.4%. This validates that compact latent tokens can effectively encode dense 3D knowledge. **+Intermediate Feature Alignment (ID #3):** Bridges the geometric-semantic gap by aligning with the 2D-3D fusion teacher, yielding an additional +1.2% gain. This confirms that raw 3D features require semantic alignment to be effectively utilized during reasoning. **+Reinforcement Learning (ID #4):** Teaches the model to actively engage spatial knowledge and reason with its 3D representations during inference, achieving the best result of 61.5%. Each stage contributes complementary improvements, validating our hypothesis that both construction and reasoning capabilities are necessary for effective spatial understanding.

4.4 RQ3: 3D Concept Distillation Strategy

A core challenge is compressing dense 3D features into a small number of latent tokens. Table 4 compares three strategies: **Average Pooling:** Directly compresses 3D features through spatial averaging. While computationally efficient, this aggressive downsampling loses fine-grained geometric details, yielding the lowest performance. **Denoising-based Distillation:** Treats latent tokens as conditions for a 3D feature denoising task, preserving critical 3D information through the reconstruction objective. **Cross-Attention (Ours):** Uses latent tokens as keys/values while visual tokens query them via cross-attention, generating equally-shaped spatial features without explicit down-sampling and compelling latent tokens to learn essential 3D concepts. Driven by empirical performance, we adopt this simple yet effective method for 3D concept distillation.

Table 5: Ablation on 3D prior integration.

ID	Method	Numerical	Multiple-Choice	**Avg.**
1	3D Feature	61.1	50.1	55.7
2	Vocabulary Entries	60.5	58.1	59.4
3	Ours	61.5	59.6	60.6

Table 6: Ablation on teacher-student architecture.

	Method	Numerical	Multiple-Choice	**Avg.**
Teacher (T1)	Concat	59.6	59.1	59.4
Teacher (T2)	Add	61.6	65.7	63.6
Student	← T2	60.0	59.5	59.8
Student	← T2+Attn-Modulate	61.5	59.6	60.6

Table 7: Ablation on latent spatial token size.

Latent Size (M)	Numerical	Multiple-Choice	**Avg.**
4	61.5	59.6	60.6
8	61.0	60.3	60.7
12	61.3	59.0	60.3

Table 8: Ablation on feature alignment layers.

ID	Method	Numerical	Multiple-Choice	**Avg.**
1	Shallow	59.5	60.8	60.2
2	Deep	60.5	59.6	60.1
3	Strided	61.5	59.6	60.6

4.5 RQ4: Integration of 3D Priors

How should distilled 3D knowledge enter the reasoning process? Table 5 compares three integration strategies: **Direct 3D Feature Injection:** Uses predicted 3D features directly as next-token embeddings. This causes training instability and performance degradation due to the semantic gap between geometric and vision-language spaces. **Vocabulary Entry Only:** Treats latent tokens as standard vocabulary items, relying solely on their embeddings. While stable, the lack of constraints on intermediate layers leaves their representations less spatially discriminative. **Vocabulary Entry + Feature Alignment (Ours):** Maintains latent tokens as vocabulary entries for stable autoregression while distilling aligned geometric representations into their hidden states. This dual design achieves both robustness and discriminative spatial reasoning capacity, leading to the best results.

4.6 RQ5: Teacher-Student Architecture

Table 6 examines architectural choices for intermediate feature distillation:

2D-3D Feature Fusion in Teacher. We compare element-wise addition versus concatenation for fusing visual and geometric features. Additive fusion achieves better feature alignment and higher downstream performance (63.6%). We therefore adopt it as the default teacher architecture.

Progressive Attention Modulation. Based on additive fusion choice, we compare distilling from a vanilla-trained teacher versus our attention-modulated variant. The modulated teacher, where answer-to-visual attention is progressively attenuated, produces latent tokens that serve as more effective information bottlenecks. Distilling from such a teacher yields +0.8% improvement, confirming that the modulation strategy forces latent tokens to aggregate well-aligned 3D information more comprehensively.

4.7 RQ6: Hyperparameter Sensitivity

Latent Spatial Token Size. Table 7 ablates the number of latent spatial tokens M. We observe that increasing M does not yield significant performance gains, indicating that a small set of latent tokens is sufficient to aggregate essential 3D knowledge and support subsequent reasoning. To maintain a compact and efficient design, we set $M = 4$ as the default configuration.

Intermediate Feature Alignment Layers. Table 8 ablates which LLM layers to distill during teacher-student intermediate feature alignment. We compare three schemes: aligning only shallow layers (the first half of the LLM), only deep layers (the second half), and our uniformly-strided alignment, which selects layers with a fixed stride of 2. In contrast to the first two schemes, our method ensures broader geometric-semantic coverage of the teacher's representations while maintaining the student's ability to adaptively integrate spatial cues, resulting in the strongest overall performance.

5 Conclusion

We have presented Latent Spatial Reasoning (LSR), a framework that enables MLLMs to construct internal 3D representations and actively reason with them during inference. LSR distills geometric knowledge into latent spatial tokens in a form compatible with the model's vision-language dynamics. Through hierarchical distillation, LSR transfers 3D structure from foundation models, bridges the

geometric–semantic gap, and trains the model to interleave spatial and linguistic reasoning, all without external 3D modules at inference time. Experiments across diverse benchmarks demonstrate the effectiveness of this approach. A potential limitation is the reliance on pretrained 3D foundation models as distillation teachers; exploring self-supervised alternatives remains a direction for future work.

Acknowledgements

This paper is supported in part by the National Natural Science Foundation of China (NO. 62322608) and in part by the Shenzhen Loop Area Institute under grant FPF10120260001. The authors would like to thank National Supercomputer Center in Guangzhou for providing high performance computational resources.

References

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.

[2] 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*, 2025.

[3] 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.

[4] 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.

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[6] 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.

[7] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26428–26438, 2024.

[8] 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*, 2025.

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[11] 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. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

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

[13] Ze Feng, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. Em-kd: Distilling efficient multimodal large language model with unbalanced vision tokens. *arXiv preprint arXiv:2511.21106*, 2025.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] 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.

[17] 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. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

[18] Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. Mllms need 3d-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025.

[19] 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.

[20] 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*, 2025.

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

[22] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

[23] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025.

[24] 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. *arXiv preprint arXiv:2510.08531*, 2025.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[28] 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.

[29] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

[30] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021.

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

[32] 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.

[33] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

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

[35] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[36] 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*, 2025.

[37] Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

[38] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[39] 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*, 2025.

[40] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[41] 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.

[42] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.

[43] LCT Xiaomi and Core Team. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 1(2): 5, 2025.

[44] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025.

[45] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[46] 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*, 2025.

[47] 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*, 2025.

[48] 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. *arXiv preprint arXiv:2505.23764*, 2025.

[49] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

[50] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

[51] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[52] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.

[53] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025.

[54] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

[55] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[56] 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.

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

[58] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[59] 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.

[60] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Implementation Details

A.1 Training Datasets

Our LSR framework is designed for robust spatial reasoning across diverse tasks. To this end, we curate a mixed training corpus by sampling from VSI-590K [47] (video-based spatial QA), SPAR-7M [52] (image-based spatial QA), and the LLaVA-Hound split [55] (general image understanding).

Video-based Spatial Reasoning. VSI-590K is primarily a video-based spatial-reasoning dataset built from real-world indoor scene videos (e.g., SCanNet [10], ScanNet++ [51]) as well as synthetic data (e.g., ProcTHOR [11] and Hypersim [30]). From this corpus, we sample approximately 240K examples for training, which are specifically sourced from four indoor scenes datasets: ScanNet [10], ScanNet++ [51], ARKitScenes [3], and S3DIS [1].

Image-based Spatial Reasoning. SPAR-7M is a large-scale dataset spanning three input modalities (i.e., single-view, multi-view, and video) and covering fine-grained tasks such as object-object and object-camera spatial reasoning. We leverage it to enhance LSR's image-based spatial understanding and sample approximately 100K examples for training.

General Image Understanding. In addition, we incorporate 64K examples from the LLaVA-Hound split to preserve foundational general image understanding and instruction-following capabilities.

A.2 Training Details

Our method is trained with standard supervised fine-tuning (SFT) followed by on-policy reinforcement learning (RL). The visual encoder and projector remain frozen throughout all stages. During 3D Concept Distillation and Intermediate Feature Alignment, we fine-tune the LLM backbone and cross-attention module with learning rate 1×10^{-5}, setting $\lambda = \beta = 1$. In the RL stage, we lower the learning rate to 1×10^{-6} and fine-tune only the LLM backbone. Below, we provide the implementation details for the RL stage.

Data for RL Training. Given computational and time budgets, we curate 10K rule-verifiable examples from our SFT dataset to form the training corpus for reinforcement learning.

Prompt Design. The user prompt employed for RL training is as follows:

User Prompt

{Question} Please think about this question as if you were a human pondering deeply. First, imagine the 3D scene layout, then reason through the answer based on this spatial context.
Requirement:
1. Represent the 3D scene layout using the special token <| lsr_pad|>, and place it between <|lsr_start|>and <|lsr_end|>tags.
2. Provide your detailed reasoning within the <think></think>tags and give your final answer between the <answer></answer>tags.

Reward Function. The reward function $R(\boldsymbol{\tau})$ is designed to assess both **answer correctness** and **format compliance** in spatial reasoning tasks. Depending on the question type, the accuracy component takes two distinct forms:

- **Multiple-Choice Questions:** For multiple-choice items, the reward is binary:

$$R_{\text{acc}}(\tau) = \mathbb{I}(Pred = GT), \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $Pred$ is the model's answer, and GT is the ground-truth choice.

- **Numerical-Answer Questions:** For numerical answers, we compute an accuracy score based on relative error tolerance. Let $Pred$ denote the predicted value and GT the target value. The normalized absolute error is defined as:

$$\text{err}(Pred, GT) = \frac{|Pred - GT|}{GT}. \tag{13}$$

The reward is then obtained by averaging over a set of confidence thresholds:

$$R_{\text{acc}}(\tau) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbb{I}\big(\text{err}(Pred, GT) \leq 1 - c\big), \tag{14}$$

where $\mathcal{C} = \{0.5, 0.55, 0.6, \ldots, 0.95\}$ is a set of uniformly spaced confidence thresholds.

A.3 Evaluation Details

For evaluation, we adopt the LMMs-Eval [53] framework. We evaluate on benchmarks spanning three categories: single-image reasoning (CVBench [35]), multi-view image reasoning (SPAR-Bench [52], MMSI-Bench [48], ViewSpatial [22]), and video-based reasoning (VSI-Bench [45]). For the video-based VSI-Bench, we uniformly sample 32 frames per video as model input.

Among them, MMSI-Bench and ViewSpatial explicitly incorporate **out-of-distribution (OOD)** evaluation protocols, designed to probe generalization under stricter and previously unseen conditions:

MMSI-Bench: A rigorous benchmark for Multi-Image Spatial Inference. It curates questions from a broad spectrum of real-world domains (including autonomous driving like nuScenes [5]/Waymo [33], robotics like AgiBot [4], and first-person everyday activities like Ego4D [14]), testing our model's generalization ability under novel domains (outdoor mixed real-world scenes).

ViewSpatial: This benchmark specifically evaluates perspective-taking ability, requiring the model to reason about spatial relationships from both egocentric (camera) and allocentric (human) viewpoints. It tests the capability of viewpoint transformation, which is not explicitly supervised in the training data, thereby evaluating our model's generalization ability under novel viewpoints (allocentric). Critically, half of the samples are drawn from the natural-image COCO [25] dataset, which further increases the OOD difficulty of this benchmark.

B More Experiments

B.1 Ablation on 3D Foundation Model

In this section, we present additional ablation studies on the 3D foundation models used in 3D concept distillation. Specifically, we introduce π^3 [40] as an alternative foundation model and compare its distillation effectiveness against VGGT.

As showcased in Table 9, experimental results demonstrate that our approach is general and robust across different 3D foundation models. Compared to the plain SFT baseline, distillation from either π^3 or VGGT yields significant gains of +2.5% and +3.3%, respectively.

Table 9: Ablation on foundation model for 3D concept distillation.

Method	Foundation Model	Numerical	Multiple-Choice	**Avg.**
Qwen2.5-VL-3B (Baseline)	-	59.5	52.6	56.1
LSR-3B (Ours)	π^3	58.4	58.7	58.6
LSR-3B (Ours)	VGGT	60.5	58.1	59.4

B.2 Comparison with Explicit Textual Reasoning

To compare latent spatial reasoning with explicit textual 3D reasoning, we start from the plain SFT baseline (Table 3, ID #1), and further train it with GRPO to reason about the 3D scene layout in

text before answering, using the same RL data and reward as LSR. As shown in Table 10, LSR outperforms this textual CoT baseline on VSI-Bench by 3.7 points while producing shorter responses on average (96 vs. 164 tokens), suggesting that latent spatial tokens provide a more compact and effective alternative to explicit textual 3D reasoning.

Table 10: Comparison with explicit textual reasoning.

Method	Avg. Performance ↑	Response Length ↓
Textual CoT	57.8	$\sim$ 164 tokens
LSR-3B (Ours)	61.5	$\sim$ 96 tokens

B.3 Efficiency Analysis

We further provide an efficiency comparison under the 32-frame input setting:

Table 11: Efficiency comparison of LSR and other approaches.

Method	External 3D Module?	Params at Inference	Peak GPU Memory (GB)
Qwen2.5-VL-3B (Baseline)	No	3B	10.9
LSR-3B (Ours)	No	3B	10.9
VG-LLM-4B [56]	Yes (VGGT)	4B	12.7

For our LSR model, the cross-attention module for 3D concept distillation and the teacher model for Intermediate Feature Alignment are employed exclusively during training. During inference, LSR model retains the same architecture as Qwen2.5-VL, except for the introduction of only 4 additional latent spatial tokens. Consequently, our approach incurs negligible additional computational overhead.

B.4 Qualitative Analysis

Figure 4 presents example analyses from VSI-Bench. Through latent-space distillation and reinforcement learning, our LSR model not only builds intrinsic 3D awareness but also actively reasons with 3D information, jointly inferring geometric relationships and contextual cues to derive the final answer. These examples illustrate how LSR moves beyond static representation to achieve dynamic, 3D-grounded reasoning.

B.5 Attention Analysis of Latent Spatial Tokens

To validate the functional role of latent spatial tokens in the reasoning process, we present two complementary visualizations: (1) Dynamic Attention Flow within individual samples, and (2) Statistical Attention Signatures across questions.

Dynamic Attention Flow within Samples: We extract the average attention weights from answer tokens to latent spatial tokens at each decoding step, visualizing the evolution of attention focus over the course of inference. This reveals how the model dynamically grounds its reasoning onto the latent spatial substrate as the answer unfolds. As illustrated in Figure 5, latent spatial tokens receive significantly higher attention during spatial reasoning, demonstrating their critical role in the model's reasoning process.

Statistical Attention Signatures across Samples: We compute the aggregated attention weights from answer tokens to latent spatial tokens (vs. raw visual tokens) across a diverse set of VSI-Bench evaluation samples, and stratify the results by question type to quantify statistical disparities in attention allocation. As illustrated in Figure 6, the model assigns higher attention to latent spatial tokens for spatial reasoning questions, while exhibiting a stronger focus on visual tokens for queries involving temporal reasoning.

C Broader Impact, Advantages, and Limitations

Broader Impact. LSR enables MLLMs to perform 3D spatial reasoning from 2D inputs without external geometric modules, advancing applications in robotics, AR/VR, and autonomous systems.

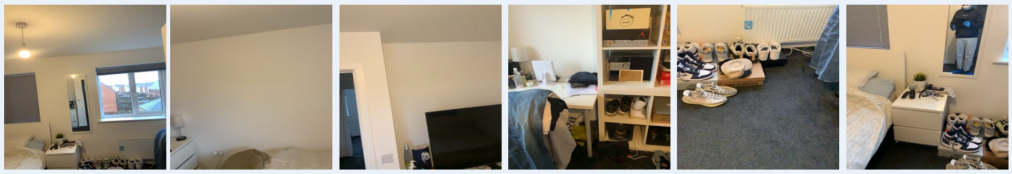


Question: Measuring from the closest point of each object, which of these objects (bed, shoes, chair, ceiling light) is the closest to the heater?

Options: *A.* bed, *B.* shoes, *C.* chair, *D.* ceiling light

Prediction: <lsr_start><lsr_pad><lsr_pad><lsr_pad><lsr_pad><lsr_end><think> Analyzing the 3D spatial layout in the video frames: The heater is positioned against the wall near the bottom-left corner of the room. The bed is placed against the wall adjacent to the heater, making it the nearest object to the heater when measuring from their closest points. The shoes and chair are further away on the floor, and the ceiling light is not in the same plane as the heater. Therefore, the bed is the closest object to the heater. </think> <answer>A</answer>,

Ground_Truth: A

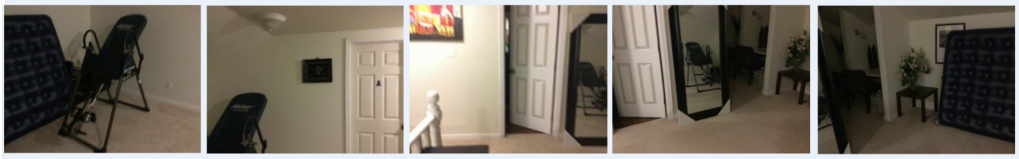


Question: If I am standing by the tv and facing the bed, is the table to the left or the right of the bed?

Options: *A.* left, *B.* right

Prediction: <lsr_start><lsr_pad><lsr_pad><lsr_pad><lsr_pad><lsr_end> <think> To determine the relative position of the table to the bed when standing by the TV and facing the bed, we need to analyze the spatial layout. The TV is positioned on the desk, and the bed is in front of it. The table is placed to the right side of the bed when facing it. This can be inferred from the perspective of someone standing by the TV and facing the bed, where the table would naturally be to the right of the bed. </think> <answer>B</answer>

Ground_Truth: B

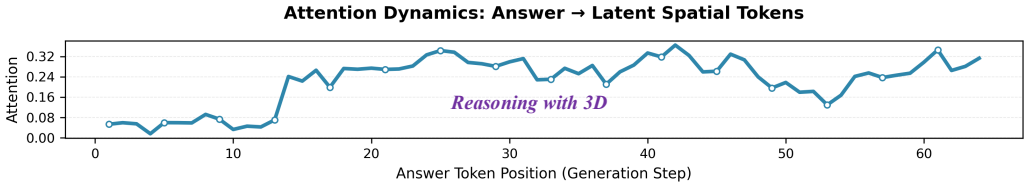


Question: Measuring from the closest point of each object, what is the distance between the mirror and the chair (in meters)?

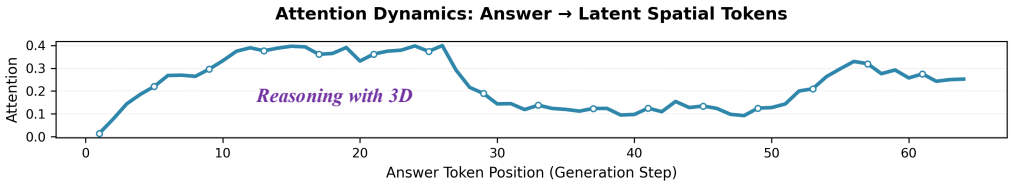
Prediction: <lsr_start><lsr_pad><lsr_pad><lsr_pad><lsr_pad><lsr_end> <think> To determine the distance between the mirror and the chair, we must consider their closest points in the 3D space. The mirror is affixed to the wall, and the chair is positioned in front of it. The closest points would be the edge of the mirror near the wall corner and the edge of the chair closest to the mirror. Observing the room's scale from the visible objects like the doorways and furniture dimensions, we can estimate the distance to be around 0.8 meters. </think> <answer>0.8</answer>

Ground_Truth: 0.9

Figure 4: Qualitative examples from VSI-Bench.



<think> To determine the direction of the refrigerator relative to the stove and table, we need to consider the spatial layout. If you're standing by the stove and facing the table, the refrigerator would be to your right side. The refrigerator is not in the front or back quadrant, but rather to the right. Therefore, the correct answer aligns with the refrigerator being to the right. </think> <answer>A</answer>



<think> The window appears rectangular, and its longest dimension seems to be its horizontal length. Comparing it to the scale of the bathroom, we can estimate the window's dimensions. Given that standard bathroom fixtures often have a proportion of about 30 inches to 40 inches for their longest dimension, we can use this as a rough scale. Scaling the window's longest dimension to approximate 35.4 inches gives an estimated length of 91.7 cm. </think> <answer> 91.7 </answer>

Figure 5: Dynamic attention flow within samples (from answer tokens to latent spatial tokens during inference).

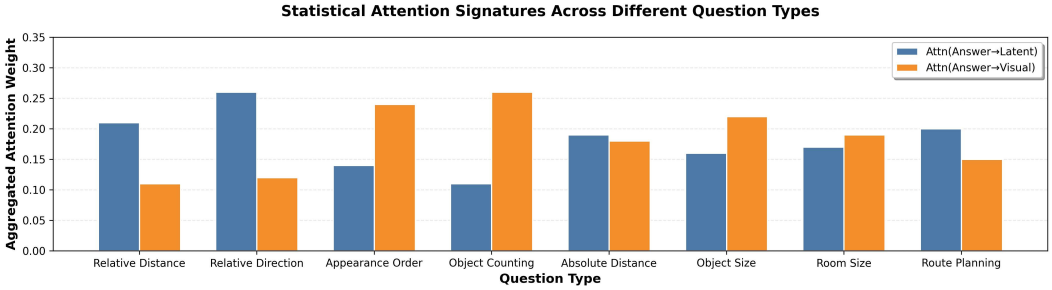


Figure 6: Statistical attention signatures across questions.

The latent-space distillation paradigm also offers a general template for injecting structured knowledge into language models. We note that enhanced spatial reasoning capabilities require responsible deployment, including attention to privacy, data consent, and transparency.

Advantages. (1) **Inference efficiency**: No external 3D encoders or depth estimators needed at test time; geometric reasoning is conducted via a small set of latent spatial tokens ($M = 4$), reducing computational overhead and enabling deployment in resource-constrained environments. (2) **Dynamic reasoning**: Latent tokens actively participate in autoregressive generation, enabling fluid spatial–linguistic interleaving. (3) **Strong generalization**: SOTA results across video (VSI-Bench), multi-view (SPAR-Bench), and single-image (CVBench) benchmarks. (4) **Modular design**: The hierarchical distillation pipeline is compatible with different MLLM backbones and 3D teachers (e.g., VGGT, π^3).

Limitations. The current framework leverages pretrained 3D foundation models to seed geometric knowledge, and extending the approach to self-supervised or task-specific teachers represents a natural direction for broader applicability. The reinforcement learning stage benefits from rule-verifiable QA pairs, suggesting that exploring more flexible reward designs could further enhance open-domain generalization. Additionally, the multi-stage training pipeline offers opportunities for future streamlining to improve accessibility across diverse computational settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims exactly reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided all the information needed to reproduce the results of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data and code will be made publicly available upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified all the training and test details necessary to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Statistical analysis was not performed due to prohibitive computational costs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided sufficient information on the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work fully adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are discussed in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all related paper in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in the paper are well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.