BEYOND GEOMETRIC ALIGNMENT: MAKING 3D LATENTS MATTER FOR SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans perform spatial reasoning by constructing internal spatial representations that support mental rotation, viewpoint transformation, and relational inference. Inspired by this ability, recent multimodal large language models introduce 3D latent reasoning, in which continuous latent tokens are expected to simulate spatial imagination steps. However, it remains unclear whether these latent tokens actually encode meaningful 3D geometry and contribute to spatial reasoning. In this work, we systematically analyze 3D latent reasoning and find that latent tokens have only weak causal influence on spatial reasoning and exhibit highly homogeneous representations across scenes. To reveal the underlying cause, we further identify an *image-dominant shortcut* during geometric alignment, whereby the alignment can be achieved largely from image features with limited reliance on 3D latents. Based on this diagnosis, we propose **CALD-3D**, a framework that combines contrastive geometric alignment with intra-latent diversity regularization to make latent tokens both geometrically meaningful and causally influential in spatial reasoning. Experiments on MindCube-Tiny and VSI-Bench demonstrate consistent improvements in spatial reasoning performance while substantially alleviating latent representation collapse.

1 INTRODUCTION

Spatial reasoning requires integrating spatial information, imagining the full scene and inferring unseen spatial relationships based on limited views, which is essential for embodied agents to understand and interact with the physical world (Yang et al., 2025b). Although Multimodal Large Language Models (MLLMs) have achieved remarkable progress in vision-language understanding, they still struggle to integrate partial observations (Xu et al., 2026) and reason about spatial relationships from alternative viewpoints (Ma et al., 2025; Shiri et al., 2024; Li et al., 2026b).

Latent visual reasoning (Li et al., 2026a; Yang et al., 2026; Wang et al., 2026b) provides a promising approach by introducing continuous latent tokens as intermediate states before textual generation. For spatial tasks, these tokens can potentially serve as a mental representation for spatial imagination and scene modeling. Following this intuition, 3DThinker (Chen et al., 2026) aligns generated 3D latent tokens with geometric features from a pretrained 3D foundation model (i.e., VGGT (Wang et al., 2025a)).

To investigate whether 3D latent tokens served as effective mental representations, we examine both their functional contribution and the geometric alignment intended to ground them. Recent analyses of 2D latent visual reasoning (Li et al., 2026e; Guo et al., 2026; Viveiros et al., 2026) have shown that generating latent tokens does not necessarily mean that their content is effectively used. In 3D latent reasoning, this concern raises a more specific question: does geometric alignment produce scene-specific 3D latent representations that causally support spatial reasoning? We address this question through three complementary analyses of latent utilization, representational diversity, and latent dependence in geometric alignment.

First, we analyze the causal role of 3D latent tokens in model predictions through inference-time interventions. Surprisingly, even severe corruption or complete removal of the latent sequence causes only marginal performance degradation, indicating that 3D latent reasoning still suffers from latent bypass despite explicit geometric supervision. This weak dependence motivates us to investigate

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.

the geometric representations learned by 3D latent tokens. By examining their cross-sample diversity and sensitivity to scene perturbations, we find that 3D latents remain highly similar and exhibit limited variation as the observed scene changes, suggesting that they fail to form distinct geometric representations. These observations lead us to further explore the geometric alignment mechanism. Existing work introduces geometric supervision by combining image embeddings and latent states for alignment with 3D geometric features, potentially allowing the alignment objective to be largely satisfied by the image branch alone. Through interventions on the two input branches separately, we reveal a strong asymmetry: perturbing the image features substantially alters the projected outputs, whereas perturbing the latent states has only a limited effect. This finding points to an *image-dominant shortcut*, whereby geometric alignment can be achieved without providing sufficient supervision for the latent states to encode geometric information.

Based on these findings, we propose **CALD-3D** (**C**ontrastive **A**lignment and **L**atent **D**iversity for 3D Latent Reasoning), a contrastive learning framework that explicitly constrains the alignment process to rely on the latent branch. For each training sample, we construct positive and negative features by projecting the same image embeddings jointly with matched latents and cross-scene latents sampled from the latent memory bank, respectively. Using 3D features extracted from the frozen VGGT as a shared anchor, the contrastive objective favors the positive projection over the negative projections, encouraging the 3D latents to encode geometry-relevant information that contributes to the alignment. In addition, we introduce *intra-latent diversity regularization*, which maintains sufficient variation across latent states while penalizing excessive pairwise similarity, preventing the latent sequence from collapsing into redundant representations.

Experiments on two representative spatial reasoning benchmarks, MindCube-Tiny (Yin et al., 2025) and VSI-Bench (Yang et al., 2025a), show that **CALD-3D** achieves consistent improvements over previous 3D spatial reasoning methods, and substantially alleviates the representation collapse problem. Our main contributions are summarized as follows:

- We provide a systematic analysis of 3D latent reasoning, revealing that 3D latents exhibit weak causal influence on model predictions and fail to form distinct scene-dependent geometric representations.
- We identify an image-dominant shortcut in the geometric supervision mechanism, where the projector relies predominantly on image features while making limited use of the latent branch, providing a plausible explanation for both latent bypass and representation collapse.
- We propose **CALD-3D**, which combines contrastive geometric grounding with intra-latent diversity regularization to improve the effectiveness of 3D latent reasoning. Extensive experiments demonstrate improved spatial reasoning performance and more diverse latent representations.

2 Related Work

2.1 Latent Visual Reasoning

Latent visual reasoning augments MLLMs with continuous latent states, allowing visual information to participate in reasoning without explicit textual generation. Existing methods mainly differ in how these latents are grounded and optimized. Mirage (Yang et al., 2026) compresses features from intermediate reasoning images, while LVR (Li et al., 2026a) aligns latent states with query-relevant visual representations. ILVR (Dong et al., 2026) introduces selective perceptual modeling, and Monet (Wang et al., 2026b) adopts teacher-state distillation with latent policy optimization. Other approaches further explore alternative supervision signals (Qin et al., 2025), spatial-semantic grounding (Cui et al., 2026), and latent optimization mechanisms (Wang et al., 2026c).

Recent studies have begun to investigate whether the learned latent states actually play their intended role during reasoning. Li et al. (2026e) investigate the causal effectiveness of latent visual imagination and propose explicit text-space imagination as an alternative. Viveiros et al. (2026) further attribute this limited dependence to weakly informative oracle supervision and latent collapse. Guo et al. (2026) show that much of the performance gain can persist even when latent slot contents are removed or altered. However, existing works focus primarily on 2D latent reasoning settings, leaving 3D latent reasoning largely unexplored. We fill this gap through a systematic analysis of the representations and functional roles of 3D latents.

2.2 Spatial Reasoning

Despite strong progress in general visual understanding, current MLLMs still exhibit substantial limitations in spatial reasoning (Wu et al., 2026b; Zhang et al., 2026a). With the emergence of 3D foundation models (Wang et al., 2025a;b; 2024), recent studies have increasingly incorporated geometric priors into MLLMs. One paradigm explicitly fuses geometric features from 3D encoders with semantic visual representations (Wu et al., 2026a; Zheng et al., 2026; Fan et al., 2026). Further refinements include adaptive geometry selection (Li et al., 2026c) and hierarchical injection of multi-level geometric features (Wang et al., 2026a; Zhang et al., 2026b). Another paradigm, exemplified by 3DRS (Huang et al., 2026), distills geometric knowledge from pretrained 3D models into MLLM visual representations. The former depends on external geometry encoders at inference, while the latter may perturb semantic representations by directly imposing geometric priors on visual features.

3DThinker (Chen et al., 2026) supervises latent states with geometric features during training, aiming to internalize 3D priors without requiring external geometry inputs at inference. Nevertheless, geometric alignment alone does not guarantee that such supervision is effectively absorbed by the latent states or that they develop sufficiently informative geometric representations. Our approach targets this limitation by strengthening the supervision of 3D latent reasoning, encouraging geometric information to be more effectively encoded into the latent states.

3 Analysis of 3D Latent Reasoning

3.1 Problem Formulation

Latent visual reasoning extends the conventional text-only autoregressive decoding by introducing continuous latent tokens as intermediate reasoning states. Given a set of input images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ and a question Q, the model can enter a *latent mode* by generating $\langle$`latent_start`$\rangle$, where the last hidden state at each step is directly fed as the next input, rather than being decoded into a discrete text token. After producing a fixed-length sequence of latent tokens $Z = \{z_k\}_{k=1}^{K}$, the model outputs $\langle$`latent_end`$\rangle$ to mark the end of latent reasoning and returns to standard text decoding. In 3D latent reasoning framework, these latent tokens are supervised with 3D geometric features during training, encouraging them to capture the underlying spatial structure across multiple views.

3.2 Do 3D Latent Tokens Causally Contribute to Spatial Reasoning?

We first investigate the role of generated latent tokens in 3D spatial reasoning process. Although latent bypass has been widely observed in 2D latent reasoning (Viveiros et al., 2026), the demands of spatial reasoning may encourage stronger reliance on latent states as intermediate spatial representations. To explore their causal influence on model predictions, we consider four interventions: replacing the latent sequence with distribution-matched Gaussian noise, zero vectors, or one generated from a different scene, and skipping latent generation by forcing $\langle$`latent_end`$\rangle$ immediately after $\langle$`latent_start`$\rangle$. The first three interventions disrupt latent content, whereas the last removes the entire latent reasoning stage. We evaluate these interventions on MindCube-Tiny (Yin et al., 2025) and VSI-Bench (Yang et al., 2025a). For MindCube-Tiny, we use the released 3DThinker checkpoint. For VSI-Bench, we reproduce the model from SpatialLadder (Li et al., 2026d) since the corresponding checkpoint and training corpus are not publicly available.

As shown in Figure 1, the overall benchmark scores of 3DThinker change little under latent intervention. On MindCube-Tiny, perturbing the latent content leads to an average accuracy drop of only 2 percentage points. More notably, model performance remains largely unchanged even when the latent mode is completely skipped, suggesting limited reliance on the generated latent tokens during subsequent reasoning. A similar pattern holds on VSI-Bench, where performance after skipping latent mode even marginally exceeds the base setting, which indicates that latent bypass persists in 3D spatial reasoning.

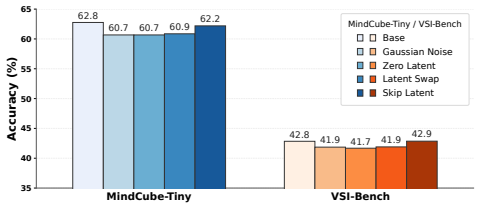

Figure 1: **Effect of latent interventions on MindCube-Tiny and VSI-Bench.** Performance changes only marginally under all interventions.

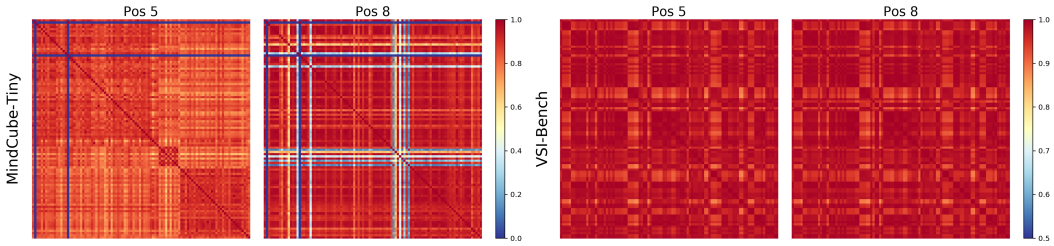

Figure 2: **Cross-sample similarity of 3D latent representations.** Pairwise cosine similarity at representative latent positions remains predominantly high on both MindCube-Tiny and VSI-Bench, indicating limited cross-sample diversity in the learned latent representations.

However, this weak dependence varies across tasks. Detailed task-level results and analysis are provided in Appendix C.1. We observe latent interventions cause larger performance degradation on **Rotation** in MindCube-Tiny and **Route Plan** in VSI-Bench, while the remaining tasks are substantially less affected. Both require more complex multi-view integration and reference-frame transformations, suggesting latent tokens may play a greater role in more demanding spatial reasoning. Nevertheless, this contribution remains limited and task-dependent, motivating stronger geometric grounding and more effective utilization of 3D latents to support spatial reasoning more reliably.

> Finding 1: Despite geometry-grounded supervision, 3D latent reasoning model exhibits weak causal dependence on latent tokens during inference.

3.3 Do 3D Latent Tokens Form Distinct Geometric Representations?

The weak latent dependence observed above leaves open whether useful geometric information is encoded but underutilized, or whether the latent representations themselves lack scene-specific content. To distinguish between these possibilities, we examine whether the learned latent representations vary across inputs or collapse toward similar patterns.

Cross-Sample Latent Diversity. We randomly sample 100 test instances from each of MindCube-Tiny and VSI-Bench, and perform inference with 3DThinker under the standard latent reasoning setting. For each instance, we extract the final-layer hidden states of the generated latent tokens, and compute pairwise cosine similarity at corresponding latent positions across instances.

As shown in Figure 2, the learned 3D latent representations exhibit consistently high inter-instance similarity on both benchmarks, as reflected by the predominance of high-valued regions in the heatmaps. Latent states remain largely invariant across most inputs, indicating limited cross-sample diversity and a pronounced tendency toward representation collapse. Notably, this homogeneity persists among samples from different scenes, which is unexpected for 3D latents intended to capture scene-specific geometric information.

Sensitivity to Scene Perturbations. To isolate the effect of scene variation, we next examine how latent states respond to controlled changes in the observed scene. Specifically, we focus on the **Among** subset of MindCube-Tiny and progressively perturb one to three input views using two scene-level intervention strategies. We then measure the resulting change in latent representations using the average cosine distance $\Delta_{\cos}$, where a larger value indicates greater deviation from the original representations. Detailed experimental settings and results are provided in Appendix C.2.

Although $\Delta_{\cos}$ increases with the number of perturbed views, latent representations remain highly similar to their unperturbed counterparts under both strategies (see Table 9). Thus, the cross-sample homogeneity is accompanied by limited responsiveness to controlled scene variations. Taken together, these observations suggest that the learned 3D latents insufficiently capture scene-specific differences, raising concerns about their ability to represent distinct scene geometry.

Finding 2: 3D latent representations show high cross-sample similarity and limited sensitivity to changes in scene structure.

3.4 Why Does Geometric Supervision Fail to Ground 3D Latent Tokens?

The weak utilization and homogeneous representations of 3D latents raise a more fundamental question: whether geometric supervision is effectively imposed on the latent states in the first place. In the 3D latent reasoning framework, as instantiated by 3DThinker, the generated latent states Z are not directly aligned with the target VGGT features F_{3D} due to the mismatch between their representation spaces. Instead, a projector P_ϕ jointly takes Z and the image embeddings H_I extracted by the vision encoder, and maps them into the geometric feature space for alignment with F_{3D}:

$$\hat{F}_{3D} = P_\phi(Z, H_I), \qquad \mathcal{L}_{\text{geo}} = \left\| \hat{F}_{3D} - F_{3D} \right\|_F^2, \tag{1}$$

This alignment objective only constrains the joint projected features but does not explicitly require the latent branch itself to encode sufficient geometric information. In particular, the image embeddings already provide rich visual cues, whereas the latent states initially lack meaningful semantics. This asymmetry makes the image branch a more direct route to predicting the target features, allowing the projector to minimize $\mathcal{L}\text{geo}$ with little contribution from Z. We refer to this behavior as an *image-dominant shortcut*:

$$P_\phi(Z, H_I) \approx P_\phi(H_I). \tag{2}$$

Branch-wise Projector Perturbation. To test the hypothesized image-dominant shortcut, we perform symmetric interventions on the latent and image branches and compare their respective contributions to geometric alignment. Following the design in Section 3.2, we consider two interventions, **Zero** and **Swap**. Specifically, while keeping one branch unchanged, we replace the other branch at the projector input with either zeros or representations from another sample.

We evaluate the effect of each intervention on MindCube-Tiny using two complementary metrics: the *relative projection change* Δ_{proj}, which measures the deviation of projected features from their unperturbed counterparts, and the *normalized alignment error increase* Δ_{align}, which quantifies the resulting degradation in matching the target geometric features. Larger values of Δ_{proj} and Δ_{align} indicate stronger branch dependence and greater alignment contribution, respectively. Detailed formulations of both metrics are provided in Appendix C.3.

As shown in Table 1, intervening on the image branch consistently causes markedly larger projection changes and alignment degradation than those on the latent branch. This strong asymmetry indicates that the projector relies predominantly on the image embeddings to achieve geometric alignment, providing direct evidence of the hypothesized image-dominant shortcut. As the alignment objective can be optimized with little dependence on the latent branch, it may provide insufficient pressure for 3D latents to encode scene-specific geometric information, helping explain their homogeneous representations and weak utilization during downstream reasoning.

Table 1: **Branch-wise interventions on the projector inputs.** The image branch dominates both projected output and geometric alignment.

Intervention	Target	Metrics	
		$\Delta_{\text{proj}} \uparrow$	$\Delta_{\text{align}} \uparrow$
Swap	Image	**0.592**	**0.230**
	Latent	0.107	0.003
Zero	Image	**1.030**	**0.908**
	Latent	0.228	0.030

Finding 3: Joint image–latent alignment admits an image-dominant shortcut, allowing geometric alignment to be achieved with limited contribution from 3D latents.

4 Method

Our analysis identifies an image-dominant shortcut in existing 3D latent reasoning, which weakens the geometric supervision received by latent states and thereby prevents them from effectively learn-

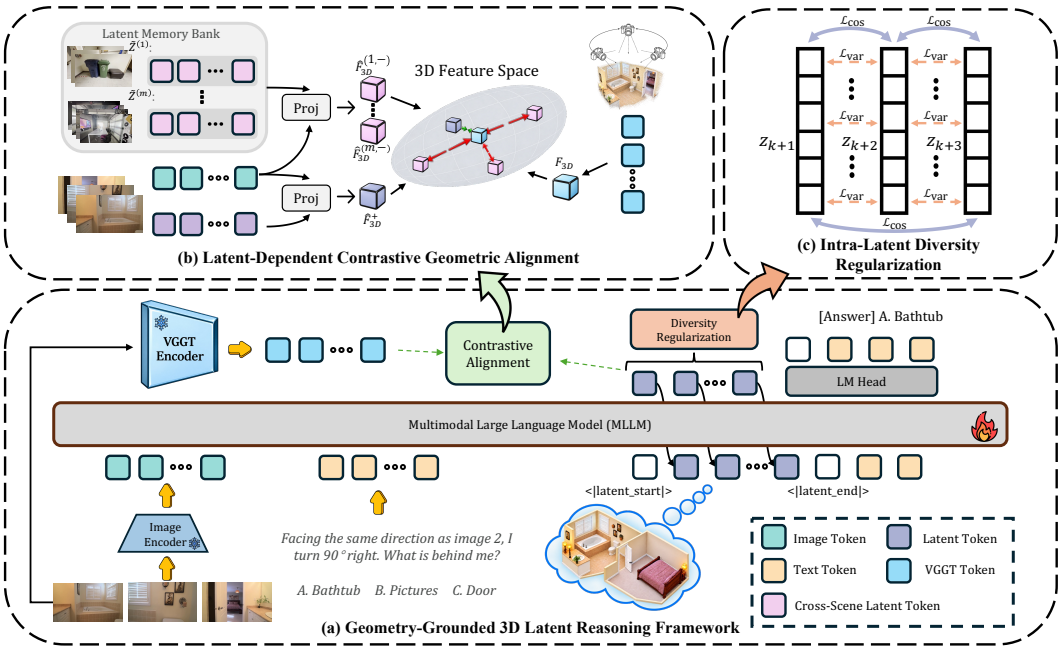

Figure 3: **Overview of CALD-3D.** (a) Geometry-grounded 3D latent reasoning framework with contrastive alignment and diversity regularization. (b) Contrasting projections from matched and cross-scene latent inputs against frozen VGGT features while keeping the image embeddings fixed. (c) Promoting intra-latent diversity through feature-wise variance and pairwise cosine constraints.

ing scene geometry. To address this issue, we propose **CALD-3D**, a framework that aims to enhance the role of 3D latents in geometric alignment while promoting diverse latent representations. As illustrated in Figure 3, CALD-3D consists of two components: (1) *Latent-Dependent Contrastive Geometric Alignment*, which contrasts matched and mismatched latent conditions to encourage greater reliance on 3D latents; and (2) *Intra-Latent Diversity Regularization*, which promotes diverse and non-redundant latent representations.

4.1 Latent-Dependent Contrastive Geometric Alignment

To explicitly encourage the alignment to leverage latent tokens rather than bypass them, we design Latent-Dependent Contrastive Geometric Alignment (Figure 3 (b)), which turns latent utilization into a discriminative signal for geometric alignment through controlled interventions on the latent branch. Specifically, for each training sample, we first retain the original projection scheme, where the generated latent states Z and the corresponding image embeddings H_I are jointly projected as $\hat{F}^{+}_{3D} = P_\phi(Z, H_I)$, which serves as a positive feature. To construct contrastive negatives, we maintain a memory bank $\mathcal{Q} = \{\tilde{Z}^{(q)}\}_{q=1}^{|\mathcal{Q}|}$ that stores latent sequences from previous training steps. We randomly retrieve M cross-scene latent sequences $\{\tilde{Z}^{(m)}\}_{m=1}^{M}$ from the memory bank $\mathcal{Q}$ and substitute them for Z, while keeping the image embeddings fixed. The resulting mismatched image-latent pairs are projected to obtain negative features $\hat{F}^{(m,-)}_{3D} = P_\phi(\tilde{Z}^{(m)}, H_I)$, where $m = 1, \ldots, M$.

Since the matched and mismatched image-latent pairs share the same image embeddings, the difference between their projected results is induced solely by the latent replacement. We then compute the cosine similarity of the positive and negative features with the anchor feature F_{3D} extracted from the frozen VGGT, respectively:

$$s^{+} = \mathrm{sim}\left(\hat{F}^{+}_{3D}, F_{3D}\right), \qquad s^{-}_{m} = \mathrm{sim}\left(\hat{F}^{(m,-)}_{3D}, F_{3D}\right), \tag{3}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Based on these scores, we formulate the contrastive loss using the InfoNCE objective (Oord et al., 2018), which encourages the positive feature to align more

closely with the geometric target than the negative features:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp\left(s^{+}/\tau\right)}{\exp\left(s^{+}/\tau\right) + \sum_{m=1}^{M} \exp\left(s_{m}^{-}/\tau\right)}, \tag{4}$$

where τ denotes the temperature parameter. Under this contrastive grounding objective, if the alignment process ignores the latent branch, latent intervention would have little effect on the projected features, resulting in only a small gap between the positive and negative scores and preventing $\mathcal{L}_{\text{con}}$ from decreasing. Minimizing $\mathcal{L}_{\text{con}}$ therefore penalizes image-dominant solutions and forces the alignment process to rely on the latent branch, encouraging the generated 3D latents to encode scene-specific geometric information.

4.2 Intra-Latent Diversity Regularization

Although contrastive alignment strengthens the contribution of the latent sequence to geometric supervision, it does not constrain how information is distributed across individual latent positions. Consequently, latent states within the same sequence may remain highly similar and encode redundant information. We therefore introduce Intra-Latent Diversity Regularization (Figure 3 (c)), which promotes feature-wise variation while reducing excessive directional similarity among latent states.

Given a latent sequence $Z = \{z_k\}_{k=1}^{K} \in \mathbb{R}^{K\times d}$, where K and d denote the sequence length and hidden dimension, respectively, we independently normalize each latent state using LayerNorm, yielding $\bar{z}_k = \mathrm{LN}(z_k)$, where $k = 1, \ldots, K$. This normalization mitigates the influence of latent scale on the diversity regularization.

Subsequently, we adapt the variance regularization of VICReg (Bardes et al., 2021) to prevent latent states from becoming homogeneous along individual feature dimensions:

$$\mathcal{L}_{\text{var}} = \frac{1}{d} \sum_{j=1}^{d} \max\left(0, \gamma - \sigma_j\right), \tag{5}$$

where γ denotes the target standard-deviation margin and σ_j denotes the standard deviation of the j-th feature dimension. This constraint penalizes insufficient variation across latent steps in each feature dimension, while imposing no additional penalty once the prescribed margin is reached.

We further introduce a complementary pairwise cosine objective to penalize excessive similarity between different latent positions:

$$\mathcal{L}_{\text{cos}} = \frac{2}{K(K-1)} \sum_{1\le i<j\le K} \max\left(0, \left|u_i^{\top} u_j\right| - \delta\right), \qquad u_k = \frac{\bar{z}_k}{\|\bar{z}_k\|_2}, \tag{6}$$

where δ denotes the margin on absolute cosine similarity. Finally, these two constraints are jointly formulated as the intra-latent regularization objective:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{var}} + \mathcal{L}_{\text{cos}}, \tag{7}$$

The variance term discourages feature-wise collapse across latent positions, while the cosine term directly reduces pairwise directional redundancy. Together, they encourage diverse and non-redundant representations within the fixed-length latent sequence.

Overall Training Objective. We augment the original geometry and text-generation losses with the proposed contrastive and regularization objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{8}$$

where the λ coefficients denote the corresponding loss weights.

5 Experiments

5.1 Experiment Settings

Training Datasets and Benchmarks. We evaluate CALD-3D on two spatial reasoning benchmarks, MindCube-Tiny (Yin et al., 2025) and VSI-Bench (Yang et al., 2025a). For MindCube-Tiny, we use

Table 2: **Main results on MindCube-Tiny**. All results are reported in accuracy (%). The best and second-best results are highlighted in **bold** and underlined, respectively. * denotes supervised training results prior to reinforcement learning.

Model	Overall	Rotation	Among	Around
Proprietary Models				
GPT-5 (Cai et al., 2026)	56.3	94.5	38.2	68.4
Gemini-2.5-Pro (Cai et al., 2026)	57.6	88.0	44.9	63.2
Base and Spatial Models				
Qwen2.5-VL-3B (Yin et al., 2025)	33.2	37.4	33.3	30.3
MindCube* (Yin et al., 2025)	60.8	47.5	62.3	67.6
SpatialSV (Tang et al., 2026)	62.3	36.0	56.2	**84.5**
Latent Models				
3DThinker* (Chen et al., 2026)	62.7	44.0	64.8	72.4
CALD-3D (Ours)	**66.9**	**58.0**	**65.8**	76.4

Table 3: **Main results on VSI-Bench.** NQ and MCQ denote numerical-answer and multiple-choice scores, respectively. * denotes supervised training results prior to reinforcement learning. † denotes our reproduced result using the same 20K training instances as CALD-3D.

Model	Data	Overall	NQ	MCQ
Qwen2.5-VL-3B (Li et al., 2026d)	–	29.4	26.0	33.0
SpatialLadder* (Li et al., 2026d)	26K	43.9	–	–
GeoThinker (Vanilla) (Li et al., 2026c)	297K	**48.9**	**56.1**	41.8
3DThinker† (Chen et al., 2026)	20K	42.8	44.7	41.0
CALD-3D (Ours)	20K	44.4	45.7	**43.1**

the publicly released training set 3DThinker-10K (Chen et al., 2026), constructed from MindCube training instances with synthesized chain-of-thought (CoT) reasoning trajectories. For VSI-Bench, we process SpatialLadder-26K (Li et al., 2026d) following a similar CoT construction protocol after excluding its object localization subset, yielding 20K training instances.

Implementation Details. We adopt Qwen2.5-VL-3B (Bai et al., 2025) as the base model and perform supervised fine-tuning. The latent sequence length is fixed to $K = 12$ during both training and inference. Additional implementation details and hyperparameters are provided in Appendix A.2.

5.2 Main Results

Table 2 reports the results on MindCube-Tiny. CALD-3D achieves the highest overall accuracy of 66.9%, surpassing the supervised 3DThinker baseline by 4.2 points and SpatialSV by 4.6 points. Compared with 3DThinker, CALD-3D improves performance across all three subtasks, with the largest gain on **Rotation**, where accuracy rises from 44.0% to 58.0% (+14.0), alongside gains of 1.0 point on **Among** and 4.0 points on **Around**. The particularly large improvement on **Rotation**, which involves viewpoint-dependent spatial transformations, is consistent with our earlier observation that latent content contributes more to demanding spatial reasoning tasks.

Table 3 reports the results on VSI-Bench. Using the same 20K training instances, CALD-3D achieves an overall score of 44.4%, outperforming 3DThinker by 1.6 points. The gains are consistent across both evaluation settings, with NQ increasing from 44.7% to 45.7% and MCQ from 41.0% to 43.1%. Notably, CALD-3D also remains competitive with GeoThinker, with an overall gap of 4.5 points despite using substantially fewer training samples (20K vs. 297K). These results demonstrate improved overall performance over 3DThinker on both benchmarks, supporting the effectiveness of CALD-3D across different spatial reasoning settings.

Table 4: **Ablation study on MindCube-Tiny**. *w/o Diversity Reg.* removes intra-latent diversity regularization; + *w/o Contrastive Align.* further removes contrastive geometric alignment; + *w/ Gaussian Noise* uses distribution-matched Gaussian noise in place of cross-scene latent states before projection to construct negative features.

Model	Overall	Rotation	Among	Around
CALD-3D (Ours)	**66.9**	**58.0**	65.8	**76.4**
w/o Diversity Reg.	65.4	49.0	**67.2**	74.4
+ w/o Contrastive Align.	61.9	48.0	63.7	68.8
+ w/ Gaussian Noise	62.3	47.5	63.8	73.6

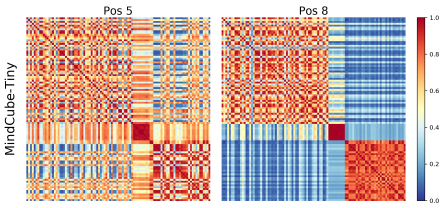

Figure 4: **Cross-sample similarity of CALD-3D on MindCube-Tiny.** Lower similarities are observed across samples.

5.3 Ablation Study

Component Analysis. We conduct progressive ablations on MindCube-Tiny to evaluate the contribution of the two components in CALD-3D. As shown in Table 4, removing the intra-latent diversity regularization (w/o Diversity Reg.) decreases the overall accuracy from 66.9% to 65.4%, confirming the benefit of explicitly encouraging diversity among latent states. Building on this variant, further removing the contrastive geometric alignment (+ w/o Contrastive Align.) reduces the accuracy to 61.9%, resulting in an additional drop of 3.5 points and a total degradation of 5.0 points from the full model. The larger performance drop highlights the importance of contrastive alignment, supporting its role in increasing the dependence of geometric alignment on latent representations.

Negative Feature Construction. We further examine how the construction of negative features affects contrastive alignment under the setting without diversity regularization. Replacing latent states with distribution-matched Gaussian noise (+ w/ Gaussian Noise) yields an overall accuracy of 62.3%, only 0.4 points above the variant without contrastive alignment (61.9%), while falling 3.1 points below the 65.4% achieved using cross-scene latent replacement for negative construction. Although Gaussian noise approximates the statistics of real latents, it does not preserve the structure of valid latent representations. This can make the resulting negatives easier to distinguish, thereby weakening the contrastive signal. These results highlight the importance of using structurally valid yet semantically mismatched latents for negative construction.

5.4 Analysis on Latent Representations

To examine whether CALD-3D alleviates the representation collapse identified in Section 3.3, we repeat the cross-sample similarity analysis on the same 100 MindCube-Tiny instances. As shown in Figure 4, CALD-3D exhibits substantially lower and more heterogeneous inter-instance similarity than 3DThinker (Figure 2), indicating increased cross-sample diversity and stronger input dependence in the learned latent representations. Moreover, at the 8th latent position, three local high-similarity clusters align with the three MindCube categories, suggesting that the learned representations capture shared structure associated with different viewpoint transformations.

We further repeat the latent intervention experiments from Section 3.2, where CALD-3D exhibits greater sensitivity to latent perturbations, providing additional evidence of increased latent utilization. Detailed results are provided in Appendix B.3.

6 Conclusion

In this work, we investigate whether geometric supervision produces informative latent representations that contribute to 3D spatial reasoning. Our analysis of 3D latent reasoning reveals weak latent utilization and highly homogeneous representations across scenes. We further identify an image-dominant shortcut that weakens geometric supervision on latent states. Motivated by these findings, we propose **CALD-3D**, combining latent-dependent contrastive alignment with intra-latent diversity regularization. Experiments on MindCube-Tiny and VSI-Bench demonstrate improved overall performance over 3DThinker, along with exhibiting more diverse latent representations and stronger reliance on latent content. These findings highlight the importance of jointly improving latent representation quality and utilization for effective 3D latent reasoning.

AI USE STATEMENT

Generative AI tools were used to generate chain-of-thought (CoT) annotations for the filtered SpatialLadder samples from the original questions, ordered images, and canonical answers, as detailed in Appendix A.1. The annotations were automatically checked for format and answer consistency; invalid outputs were regenerated, and persistent failures were marked as null. Beyond dataset construction, generative AI was used only for language polishing. It was not used for research ideation, methodology or experimental design, method implementation, mathematical claims or proofs, result interpretation, or any other research task. All AI-assisted content was reviewed by the authors, who take responsibility for the final dataset and all reported results.

REFERENCES

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.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026.

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

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.

Jin Cui, Xinyue Long, Xunyong Zhang, Yadong Zhang, Chuanchang Su, Jingye Gan, Boran Zhao, and Pengju Ren. Retrieve, integrate, and synthesize: Spatial-semantic grounded latent visual reasoning. *arXiv preprint arXiv:2605.07106*, 2026.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved latent visual reasoning with selective perceptual modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29316–29335, 2026.

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

Jiawei Guo, Yu Chen, Xiang Wang, Shuai Li, Xinpei Zhao, Huaxing Liu, Shuai Dong, Feifei Zhai, and Yu Zhou. Beyond visual memory: Mechanistic diagnostics of latent visual reasoning. *arXiv preprint arXiv:2606.01287*, 2026.

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

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

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. In *European Conference on Computer Vision*, pp. 95–111. Springer, 2026b.

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

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

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space. *arXiv preprint arXiv:2602.22766*, 2026e.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Introducing GPT-5. `https://openai.com/index/introducing-gpt-5/`, 2025. Accessed: 2026-09-24.

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.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21440–21455, 2024.

Jiayu Tang, Yuchen Zhou, and Chao Gou. Spatialsv: Internalizing interpretable 3d spatial awareness in mllms via task-oriented visual supervision. In *IJCAI*, pp. 1685–1693, 2026.

André G Viveiros, Nuno Gonçalves, André FT Martins, and Matthias Lindemann. What's holding back latent visual reasoning? *arXiv preprint arXiv:2605.18445*, 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.

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 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 image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. Forest before trees: Latent superposition for efficient visual reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11272–11288, 2026c.

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

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31029–31041, 2026b.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Matt Feiszli, and Kevin J Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31078–31088, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025a.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025b.

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

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

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

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38678–38688, 2026b.

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

A Implementation Details

A.1 Construction of SpatialLadder CoT Data

SpatialLadder-26K (Li et al., 2026d) contains 26,610 samples spanning object localization, single-image, multi-view, and video-based spatial reasoning tasks. We exclude its approximately 5.9K object-localization samples, which focus on bounding-box prediction and differ from the reasoning-oriented tasks considered in our downstream benchmarks. We then synthesize CoT reasoning trajectories for the remaining roughly 20K samples.

Following the CoT construction protocol of 3DThinker (Chen et al., 2026), we use GPT-4.1 as the teacher model for reasoning annotation. For multiple-choice questions, all candidate options are appended to the question, with the answer label converted to its full option text. The teacher is provided with the question, ordered images, and canonical answer, and instructed to first consolidate the observed views into a mental 3D scene before generating the reasoning trajectory. The system prompt is shown in Figure 5.

We additionally assess the quality of each generated CoT. A valid output must begin with exactly one `<output_3D>` marker, followed by a non-empty textual rationale in the format `<think>...</think> <answer>...</answer>`. The content within `<answer>` must exactly match the canonical answer. We reject outputs containing a closing `</output_3D>` marker, an incorrect answer, or any text outside the prescribed structure. Invalid outputs are regenerated up to three times, while persistent failures are marked as `null`.

Core system instruction

You are an advanced AI reasoning assistant with expertise in 3D spatial understanding. Given the Question, Images, and Answer, mentally construct a 3D layout of the scene and complete the intermediate reasoning process. Begin with exactly one `<output_3D>` token, place the reasoning inside `<think>...</think>`, and place the final result inside `<answer>...</answer>`. Do not include `</output_3D>`. The final answer must exactly match the provided Answer, without mentioning that it was given or that the reasoning confirms it.

Figure 5: System instruction used to synthesize SpatialLadder CoT trajectories.

A.2 Details of Training Implementation

All experiments are conducted using Qwen2.5-VL-3B (Bai et al., 2025) as the base model. Due to computational constraints, we do not further scale the experiments to the 7B variant. Supervised fine-tuning is implemented with the Hugging Face TRL framework. During training, we freeze the vision encoder and jointly optimize the language-model backbone and the geometric projector.

The latent sequence length is fixed to $K = 12$ during both training and inference. For contrastive alignment, we sample $M = 4$ negative latent sequences per training sample and set the temperature to $\tau = 0.1$. For diversity regularization, we set the variance margin to $\gamma = 0.2$ and the cosine similarity margin to $\delta = 0.9$. The loss weights $\lambda_{\text{CE}}, \lambda_{\text{geo}}, \lambda_{\text{con}}$, and λ_{reg} are set to 0.1, 0.5, 0.3, and 0.1, respectively.

Since MindCube and SpatialLadder differ in both scale and data composition, we adopt slightly different optimization settings for the two datasets. The complete training hyperparameters are summarized in Table 5.

B Additional Experiment Details

B.1 Details of Benchmarks

We evaluate CALD-3D on two spatial reasoning benchmarks, MindCube-Tiny (Yin et al., 2025) and VSI-Bench (Yang et al., 2025a).

Table 5: Training hyperparameters for MindCube and SpatialLadder.

Hyperparameter	MindCube	SpatialLadder
per_device_train_batch_size	1	1
learning_rate	1e-4	7e-5
num_train_epochs	8	6
gradient_accumulation_steps	1	4
warmup_ratio	0.03	0.03
weight_decay	0.01	0.01
data_seed	42	42
gradient_checkpointing	true	true
bf16	true	true

MindCube-Tiny. MindCube-Tiny is a diagnostic subset of MindCube designed to evaluate spatial reasoning from limited multi-view observations, which contains 1,050 questions, including 200 **Rotation**, 600 **Among**, and 250 **Around** instances. Rotation involves reasoning across orthogonal views captured from a fixed observation point, Among requires maintaining spatial consistency across views under visibility constraints, and Around emphasizes reasoning under occlusion and changing viewpoints. We report accuracy for each category and the overall benchmark.

VSI-Bench. VSI-Bench is a video-based spatial reasoning benchmark which contains over 5,000 question-answer pairs derived from 288 egocentric indoor videos. It covers eight tasks, with four numerical-answer tasks: **Object Count**, **Absolute Distance**, **Object Size**, and **Room Size**; and four multiple-choice tasks: **Relative Distance**, **Relative Direction**, **Route Plan**, and **Appearance Order**. These tasks span configurational reasoning, measurement estimation, and spatiotemporal reasoning. Following the official evaluation protocol, numerical-answer questions are evaluated using Mean Relative Accuracy (MRA), while multiple-choice questions are evaluated using accuracy.

B.2 Details of Baselines

We compare CALD-3D with both general-purpose MLLMs and recent spatial reasoning methods. For general-purpose models, we report GPT-5 (OpenAI, 2025) and Gemini-2.5-Pro (Comanici et al., 2025) as proprietary references on MindCube-Tiny, together with Qwen2.5-VL-3B (Bai et al., 2025), which also serves as the backbone of CALD-3D.

On MindCube-Tiny, we additionally compare with the supervised MindCube model (Yin et al., 2025), which supports spatial reasoning through cognitive-map generation and free-form reasoning, and SpatialSV (Tang et al., 2026), which enhances spatial representations through explicit geometric supervision.

On VSI-Bench, we include SpatialLadder (Li et al., 2026d), which progressively develops spatial perception and reasoning capabilities, and GeoThinker (Li et al., 2026c), which selectively incorporates task-relevant geometric features during reasoning. We further use 3DThinker (Chen et al., 2026) as our primary baseline on both benchmarks. As the closest comparison to CALD-3D, 3DThinker aligns intermediate 3D latent states with geometric features from a pretrained 3D foundation model.

B.3 Latent Intervention Results of CALD-3D

Table 6 reports the latent intervention results of CALD-3D on MindCube-Tiny. Unlike the weak sensitivity observed for 3DThinker in Section 3.2, CALD-3D exhibits substantial performance degradation under all interventions. The overall accuracy drops by 6.4–7.7 points when the latent content is corrupted or removed, with **Skip Latent** causing the largest decrease from 66.9% to 59.2%. The effect is particularly pronounced on **Rotation**, where performance decreases by 15.5-19.0 points across the four interventions. These results provide further evidence that CALD-3D relies more strongly on the generated latent representations during spatial reasoning, especially for tasks involving viewpoint transformations.

Table 6: Latent intervention results of CALD-3D on MindCube-Tiny.

Model	Overall	Rotation	Among	Around
CALD-3D	**66.9**	**58.0**	**65.8**	**76.4**
w/ Gaussian Noise	59.5	39.0	60.8	<u>72.8</u>
w/ Zero Latent	60.2	40.0	61.8	72.4
w/ Latent Swap	<u>60.5</u>	<u>42.5</u>	<u>62.7</u>	69.6
w/ Skip Latent	59.2	39.5	60.3	72.4

Table 7: Detailed results of latent interventions on MindCube-Tiny.

Model	Overall	Rotation	Among	Around
3DThinker	**62.8**	**43.0**	<u>65.7</u>	<u>71.6</u>
w/ Gaussian Noise	60.7	<u>39.0</u>	63.3	<u>71.6</u>
w/ Zero Latent	60.7	37.5	64.3	70.4
w/ Latent Swap	60.9	37.0	65.0	70.0
w/ Skip Latent	<u>62.2</u>	36.5	**66.3**	**72.8**

C Additional Analysis Details

C.1 Detailed Analysis of Latent Interventions

We further report the subtask-level results of latent interventions on MindCube-Tiny and VSI-Bench in Tables 7 and 8, respectively, where **Base** denotes evaluation of the official checkpoint in our local environment without latent intervention. Although the overall performance is only mildly affected, the impact of latent intervention varies substantially across different spatial reasoning tasks.

On MindCube-Tiny, **Rotation** is the most sensitive subtask, with accuracy dropping from 43.0 to 36.5–39.0 under different interventions, while **Around** and **Among** remain comparatively stable. A similar task-dependent pattern appears on VSI-Bench, where **Route Plan** consistently degrades under all interventions, dropping from 30.4 to as low as 25.8, while the remaining subtasks exhibit more heterogeneous responses.

These results suggest that latent states are more useful for tasks requiring stronger multi-view integration and reference-frame transformation, but their contribution remains limited for most spatial reasoning tasks. This task-dependent sensitivity further indicates that geometry-grounded supervision does not lead to consistent latent utilization during inference.

C.2 Detailed Analysis of Scene Perturbations

Perturbation Setup. We conduct the scene perturbation analysis on the **Among** subset of MindCube-Tiny, where four orthogonal views provide complementary spatial evidence for inferring the overall scene configuration. This multi-view structure allows us to systematically alter partial scene evidence and examine the sensitivity of the resulting latent representations. We randomly sample five instances from each unique scene, resulting in 230 test instances in total.

Given an original four-view input $\mathcal{I} = \{I_j\}_{j=1}^{4}$, we progressively perturb one to three views. For each instance, we construct nested subsets $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \mathcal{S}_3$ with $|\mathcal{S}_s| = s$, where $s \in \{1, 2, 3\}$. Thus, increasing the perturbation level preserves the previously perturbed views while introducing additional changes.

We consider two perturbation strategies. **Mask** removes the visual content of selected views, thereby reducing the available scene evidence. **Replace** substitutes images with those from the corresponding viewpoints in another scene, introducing inconsistent evidence that disrupts cross-view coher-

Table 8: Detailed results of latent interventions on VSI-Bench. † denotes our reproduced result using the same 20K training instances as CALD-3D.

Model	Overall	Numerical Question				Multiple-choice Question			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
3DThinker†	<u>42.8</u>	**60.2**	<u>30.4</u>	<u>55.9</u>	32.2	**43.4**	41.4	**30.4**	<u>48.9</u>
w/ Gaussian Noise	41.9	58.1	27.4	52.8	<u>35.2</u>	<u>42.8</u>	**44.0**	26.8	47.7
w/ Zero Latent	41.7	<u>58.2</u>	28.5	54.7	34.0	42.3	40.5	<u>29.4</u>	46.0
w/ Latent Swap	41.9	57.0	29.7	**56.1**	32.7	<u>42.8</u>	41.9	26.8	45.5
w/ Skip Latent	**42.9**	56.5	**31.7**	54.8	**39.2**	41.8	<u>43.8</u>	25.8	**49.4**

Table 9: Average latent representation change under controlled scene perturbations on the Among subset of MindCube-Tiny.

# Perturbed Views	$\Delta_{\cos}$	
	Mask	Replace
1	0.095	0.113
2	0.134	0.155
3	**0.154**	**0.184**

ence. Formally, for perturbation type $p \in \{\text{mask}, \text{replace}\}$, the perturbed view is defined as:

$$I_j^{(p,s)} = \begin{cases} I_j, & j \notin \mathcal{S}_s, \\ \text{Mask}(I_j), & j \in \mathcal{S}_s,\ p = \text{mask}, \\ \widetilde{I}_j, & j \in \mathcal{S}_s,\ p = \text{replace}, \end{cases} \tag{9}$$

where $\widetilde{I}_j$ denotes the corresponding view from an unrelated scene.

Evaluation Metric. For each original input, we extract the $K = 12$ generated latent states, denoted by $Z = \{z_k\}_{k=1}^K$. We then perform inference on the perturbed input $\mathcal{I}^{(p,s)} = \{I_j^{(p,s)}\}_{j=1}^4$ and obtain $Z^{(p,s)} = \{z_k^{(p,s)}\}_{k=1}^K$. For each perturbation setting, we measure the change between the original and perturbed latent representations by averaging the cosine distance over all test instances and corresponding latent positions:

$$\Delta_{\cos}^{(p,s)} = 1 - \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \text{sim}\left(z_{n,k}, z_{n,k}^{(p,s)}\right), \tag{10}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and N is the number of test instances. A larger $\Delta_{\cos}$ indicates a greater deviation from the latent representations generated from the original input. If the latent representations are sensitive to scene-specific visual evidence, perturbing more views should induce progressively larger representational changes.

Results Analysis. As shown in Table 9, $\Delta_{\cos}$ increases as more views are perturbed under both strategies, indicating that the latent representations do respond to changes in scene evidence. Replace consistently induces larger changes than Mask, as it introduces cross-scene inconsistent evidence rather than simply removing visual information.

Nevertheless, the overall representational changes remain limited. Even when three of the four views are perturbed, $\Delta_{\cos}$ reaches only 0.154 for Mask and 0.184 for Replace, corresponding to average cosine similarities of 0.846 and 0.816 with the original latent states. Combined with the high cross-sample similarity observed in Section 3.3, these results suggest that the learned 3D latents remain only weakly sensitive to scene-specific variations.

C.3 Detailed Analysis of Branch-wise Projector Perturbations

Perturbation Setup. We evaluate the dependence of the geometric projector on its two input branches by independently perturbing the latent states Z and the image embeddings H_I. For each

sample i, the original projected geometric features are given by:

$$\hat{F}_{3D}^{(i)} = P_\phi\left(Z^{(i)}, H_I^{(i)}\right). \tag{11}$$

For each branch $b \in \{\text{latent}, \text{image}\}$, we apply either **Zero** or **Swap** perturbation while keeping the other branch unchanged. Zero replaces the selected branch with zero features, whereas Swap replaces it with the corresponding representation from another sample depicting a different scene. We denote the resulting projected features by $\widetilde{F}_{b,p}^{(i)}$, where $p \in \{\text{zero}, \text{swap}\}$.

Evaluation Metrics. We first quantify how strongly perturbing each branch changes the projector output using the relative projection change:

$$\Delta_{\text{proj}}^{(b,p)} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left\|\widetilde{F}_{b,p}^{(i)} - \hat{F}_{3D}^{(i)}\right\|_F}{\left\|\hat{F}_{3D}^{(i)}\right\|_F + \epsilon}, \tag{12}$$

where N denotes the number of evaluated samples and ϵ is a small constant for numerical stability. A larger Δ_{proj} indicates that the projected representation is more sensitive to the corresponding input branch.

Projection sensitivity alone, however, does not determine whether the perturbed branch is important for matching the target geometric features. We therefore additionally measure the normalized increase in geometric alignment error:

$$\Delta_{\text{align}}^{(b,p)} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left\|\widetilde{F}_{b,p}^{(i)} - F_{3D}^{(i)}\right\|_F^2 - \left\|\hat{F}_{3D}^{(i)} - F_{3D}^{(i)}\right\|_F^2}{\left\|F_{3D}^{(i)}\right\|_F^2 + \epsilon}. \tag{13}$$

Here, a larger positive Δ_{align} indicates that perturbing the branch causes a greater degradation in matching the target VGGT features, suggesting a larger contribution of that branch to geometric alignment. The two metrics are complementary: Δ_{proj} measures the sensitivity of the projector output, whereas Δ_{align} measures the corresponding effect on the geometric supervision objective.

Results Analysis. As shown in Table 1, both Swap and Zero produce substantially larger projection changes and alignment degradation when applied to the image branch than to the latent branch. The consistent asymmetry across two metrics indicates that the image branch has a markedly stronger influence on both the projector output and geometric alignment than the latent branch.