GEO-REPA: GEOMETRY-AWARE REPRESENTATION ALIGNMENT ENHANCES NATIVE SPATIAL INTELLIGENCE IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal Large Language Models (MLLMs) excel at semantic understanding but often struggle to reason about physical size, depth, and spatial relationships in 3D environments. Visual geometry encoders provide strong geometric priors, yet our analysis of frozen checkpoints on held-out samples finds that correspondence between MLLM hidden states and geometry features weakens in later layers after input-level fusion, which lacks direct intermediate supervision. We introduce **GEO-REPA**, a geometry-aware representation alignment framework that encourages intermediate MLLM features to preserve and use 3D cues. To account for camera motion between image and 3D scene coordinate frames, GEO-REPA introduces a pose-aware SE(3)-equivariant inductive bias. Learned pose tokens predict per-frame rotations and translations, which induce a transformation of intermediate visual tokens. A dual alignment objective supervises the transformed visual tokens with 3D geometry features and the pose tokens with camera pose targets from a visual geometry encoder. Across VSI-Bench, SQA3D, and Scan2Cap, GEO-REPA achieves competitive or leading spatial performance, with Scan2Cap gains extending to a newer MLLM backbone. Representation diagnostics associate stronger alignment with lower depth error, while cross-view visualizations show reduced separation among views after transformation. Quantitative analyses associate alignment with spatial performance, show pose sensitivity and meaningful rotation and translation direction estimates. Matched ablations favor pose-aware alignment over no alignment and direct matching across three tasks. On general multimodal benchmarks, the evaluated variant scores modestly below its original backbone. Neither the external geometry encoder nor the alignment modules are needed at inference. Code, scripts, and checkpoints will be released.

1 INTRODUCTION

Spatial intelligence Yang et al. (2025a) defines the capacity to perceive, reason about, and interact with 3D environments, and is essential for embodied AI applications. While Multi-modal Large Language Models (MLLMs) (Chen et al., 2024c; Bai et al., 2025; Li et al., 2025a) excel in semantic understanding, their ability to perform precise spatial reasoning, such as estimating physical size, depth, and navigational relationships, remains limited. State-of-the-art visual geometry foundation models (Wang et al., 2025b; Huang et al., 2025a; Wang et al., 2025c) show strong spatial priors by capturing inter-frame correspondences and scene structure. However, when these geometry experts are integrated into MLLM pipelines, their 3D cues may not persist in later hidden representations, so the model may fail to use them effectively for spatial reasoning.

Existing approaches to equip MLLMs with spatial capabilities largely follow two paradigms, as illustrated in Figure 1. The first line, termed **3D/2.5D-Input Models** (Figure 1a), augments visual inputs with explicit geometric signals (Zheng et al., 2025b), such as depth maps or point clouds. While effective, their reliance on auxiliary 3D or 2.5D inputs restricts their applicability in standard RGB-only scenarios. The second paradigm (Wu et al., 2025a; Zheng et al., 2025a), **Input-level Fusion** (Figure 1b), injects geometry priors by fusing visual geometry features with MLLM tokens at the input stage. However, input-level fusion alone provides no direct constraint on how these 3D

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

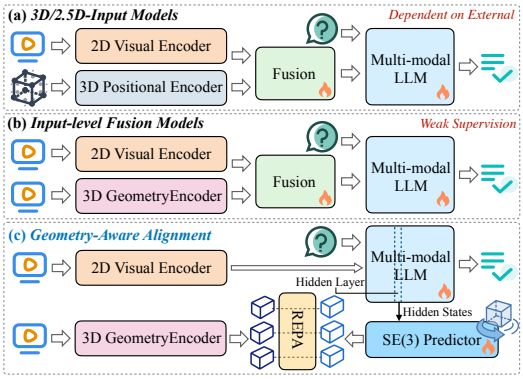
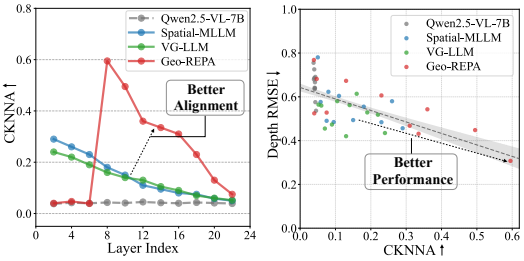


Figure 1: **Spatial Intelligence Paradigms.** (a) *3D/2.5D-input models* use auxiliary geometric inputs. (b) *Input-level fusion* injects 3D priors at the input stage. (c) *Geometry-aware alignment (ours)* aligns hidden representations with 3D features via pose-aware SE(3) transformation.



(a) Alignment to VGGT. (b) Performance vs. alignment.

Figure 2: **Representation alignment in MLLMs.** (a) CKNNA shows input-level fusion models align less with geometry features in later layers, while GEO-REPA improves alignment through intermediate supervision. (b) Stronger alignment correlates with lower Depth RMSE, suggesting improved 3D cue preservation.

cues are maintained in intermediate and later representations. As a result, these models may still underuse the injected 3D cues, limiting their performance on tasks that require spatial reasoning.

We analyze intermediate representations to examine whether input-level fusion preserves 3D cues across the MLLM backbone. In Figure 2a, we use Centered Kernel Alignment (CKNNA) (Li et al., 2025c) to measure the similarity between MLLM hidden states and geometric features across layers. CKNNA is only a diagnostic tool, not an evaluation metric. The results show modest alignment in early layers, which weakens in later layers. This raises a question: *How can MLLMs preserve and use 3D cues beyond the input stage?* We draw on **REPA** (Representation Alignment) (Yu et al., 2024), which aligns intermediate representations with strong external features to improve representation learning. However, direct feature alignment is insufficient for spatial intelligence, since MLLM visual tokens are tied to the image coordinate frame, whereas geometry encoder features are organized in a 3D scene coordinate frame. We therefore make the alignment pose-aware, using SE(3) transformations (Fuchs et al., 2020; Yim et al., 2023) to account for camera motion.

Motivated by this, we propose **GEO-REPA**, a framework for **Geo**metry-aware **REP**resentation **A**lignment that strengthens MLLM spatial reasoning by aligning intermediate representations with 3D geometric supervision. Rather than using geometry only as an external input, GEO-REPA encourages the MLLM backbone to preserve and use 3D cues in its hidden representations. Specifically, we introduce a pose-aware SE(3)-equivariant inductive bias (Yim et al., 2023; Fuchs et al., 2020) into the alignment process. Given a sequence of frames, GEO-REPA predicts per-frame rigid body motions, including rotations and translations, from learned pose tokens, and applies the induced SE(3)-equivariant transformation to intermediate visual tokens. The transformed visual tokens are then aligned with 3D geometry representations from a visual geometry encoder, e.g., VGGT (Wang et al., 2025b), while the pose tokens are aligned with camera pose targets. In this way, the alignment accounts for camera motion and coordinate-frame changes, rather than only matching visual and geometry features in a static embedding space.

Diagnostics on frozen checkpoints show greater similarity between intermediate MLLM hidden states and geometry features (Figure 2a), with stronger alignment associated with lower depth estimation error (Figure 2b). Across VSI-Bench (Yang et al., 2025b), SQA3D (Ma et al., 2022), and Scan2Cap (Chen et al., 2021), GEO-REPA achieves competitive or leading results, with captioning gains extending to another backbone. Visualizations show reduced pose-dependent separation after transformation, while alignment correlates with SQA3D performance. Pose perturbations decrease alignment, while ScanNet comparisons find lower rotation and translation direction errors than shuffled poses. Matched ablations across three tasks favor pose-aware SE(3) alignment over no alignment and direct feature matching. General multimodal scores decline modestly relative to the original backbone, while inference needs no 3D encoder or alignment heads.

- **Geometry-aware REPA.** We show that input-level fusion alone does not explicitly encourage 3D cues to be preserved in later MLLM representations. To address this, we propose GEO-REPA,

which aligns intermediate MLLM representations with 3D geometric priors during training, encouraging the backbone to preserve and use 3D cues for spatial reasoning beyond the input stage.
- **SE(3)-equivariant inductive bias in alignment.** MLLM visual tokens are tied to the image coordinate frame, whereas geometry encoder features follow a 3D scene frame, making direct matching suboptimal. We introduce an SE(3)-equivariant transformation into the alignment process. By predicting frame-wise rigid motions from learned pose tokens, GEO-REPA builds a pose-aware bridge between the two frames and accounts for camera motion.
- **Performance and analyses.** GEO-REPA achieves competitive or leading results on three benchmarks without a 3D encoder at inference. Visualizations show reduced cross-view separation, alignment correlates with SQA3D performance, and quantitative pose tests show geometric sensitivity. Matched ablations across three tasks favor pose-aware SE(3) alignment over no alignment and direct matching.

2 Related Work

MLLMs for Spatial Intelligence. MLLM spatial reasoning lags semantic understanding because language-oriented training underemphasizes fine-grained 3D cues. Spatial-MLLM (Wu et al., 2025a) and VG-LLM (Zheng et al., 2025a) fuse semantics with implicit geometry from depth-aware foundation models (*e.g.,* VGGT (Wang et al., 2025b)), without explicit inference-time 3D inputs. Unified 3D-2D models: SAT-LLaVA (Ray et al., 2024) uses auxiliary depth estimation; LLaVA-3D (Chen et al., 2024a) embeds 3D positional encodings in 2D patches. GS-Reasoner (Chen et al., 2026a), SpaceR (Ouyang et al., 2025), and 3DThinker (Chen et al., 2026b) use chain-of-thought or reinforcement learning. 3DRS (Huang et al., 2025b) directly aligns MLLM visual features with 3D foundation model features. GEO-REPA instead applies a pose-aware SE(3)-equivariant transformation before alignment, bridging image and 3D scene frames. Most use input/output geometry or direct feature supervision without explicitly modeling camera motion during alignment.

Spatial Representation Alignment. Alignment matches intermediate features to external representations. REPA (Yu et al., 2024) aligns diffusion latents with strong visual encoders (*e.g.,* DINOv2 (Oquab et al., 2023)) for efficient training. Point-SRA (Wei et al., 2026) studies local-global point-cloud structure; Sereyjol-Garros et al. (2026) align RGB priors with LiDAR synthesis; Geometry Forcing (Wu et al., 2025b) aligns video diffusion features with 3D representations for world modeling; Spatial Forcing (Li et al., 2025b) aligns vision-language-action models with spatial constraints for manipulation. Unlike these methods, GEO-REPA targets MLLM spatial reasoning, bridging the coordinate-frame mismatch between image visual tokens and 3D geometry encoder features with a pose-aware SE(3)-equivariant bias that accounts for camera motion.

3 Method

3.1 Overview of Geo-REPA

Overall. As shown in Figure 3, given a question and a frame sequence $\{\boldsymbol{I}_i\}_{i=1}^{S}$, we extract intermediate MLLM hidden states and obtain learnable pose tokens $\boldsymbol{r}_{\text{pose}}$ (initialized and optimized during training) and visual tokens $\boldsymbol{r}_{\text{2d}}$. In parallel, a Visual Geometry Encoder processes these frames and produces pose representations $\hat{\boldsymbol{r}}_{\text{pose}}$ and 3D geometry representations $\boldsymbol{r}_{\text{3d}}$. From $\boldsymbol{r}_{\text{pose}}$, an SE(3) predictor estimates per-frame rotation and translation parameters $\{\boldsymbol{R}_i\}_{i=1}^{S}$ and $\{\boldsymbol{t}_i\}_{i=1}^{S}$, which we use to apply an SE(3)-equivariant transformation $\rho(\boldsymbol{R}_i, \boldsymbol{t}_i)$ to the visual tokens to yield an SE(3)-aware representation. We then perform geometry-aware alignment between the transformed visual representations $\boldsymbol{r}_{\text{se(3)}}$ and $\boldsymbol{r}_{\text{3d}}$, while enforcing pose-level alignment between $\boldsymbol{r}_{\text{pose}}$ and $\hat{\boldsymbol{r}}_{\text{pose}}$.

Backbone and Geometry Foundation Model. To leverage the reasoning capability of MLLM backbone, we adopt a chain-of-thought prompting scheme, where the model generates a reasoning trace before producing task outputs via standard next-token prediction. Formally, given visual inputs $\{\boldsymbol{I}_i\}_{i=1}^{S}$ and a textual query Q, the policy $\pi_{\boldsymbol{\theta}}$ generates a sequence as:

$$y_t \sim \pi_{\boldsymbol{\theta}}\big(\cdot \mid \boldsymbol{y}_{0:t-1},\, Q,\, \{\boldsymbol{I}_i\}_{i=1}^{S}\big),\ t = 1, \ldots, T, \tag{1}$$

We extract $\boldsymbol{r}_{\text{pose}}$ and $\boldsymbol{r}_{\text{2d}}$ from the MLLM hidden states at an intermediate layer L, and use them as the representations to be aligned in our geometry-aware alignment module. We use a 3D visual

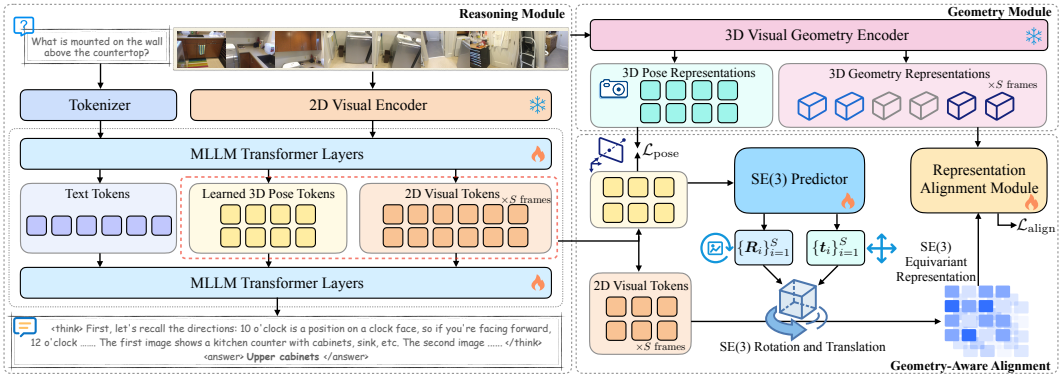


Figure 3: **GEO-REPA Overview.** Our model consists of a 2D visual encoder, an MLLM backbone with learnable 3D pose tokens, a 3D visual geometry encoder, an SE(3) predictor, and a representation alignment module. During training, we enforce SE(3)-equivariant representation alignment between visual representations and 3D geometry representations. At inference, the 3D visual geometry encoder is not required, improving inference efficiency.

geometry encoder as our geometry foundation to model inter-frame correspondences and 3D structure from the input frames. We adopt VGGT for this role, since it is trained to recover 3D scene geometry from images without relying on explicit 3D priors, making it a suitable source of geometric representations. Given $\{\boldsymbol{I}_i\}_{i=1}^S$, the VGGT encoder outputs pose representations $\hat{\boldsymbol{r}}_{\text{pose}}$ and 3D geometry representations $\boldsymbol{r}_{\text{3d}}$ for each input frame, used for subsequent geometry-aware alignment.

Geometry-Aware Alignment. We align MLLM representations with 3D geometry by imposing an explicit SE(3) structure. Specifically, an SE(3) predictor takes the pose tokens $\boldsymbol{r}_{\text{pose}}$ and estimates per-frame 3D rotation parameters $\{\boldsymbol{R}_i\}_{i=1}^S$ and translation parameters $\{\boldsymbol{t}_i\}_{i=1}^S$, where each $\boldsymbol{R}_i \in \mathbb{R}^{3\times3}$ is obtained from a continuous 6D rotation representation via orthogonalization and $\boldsymbol{t}_i \in \mathbb{R}^3$. We then apply the corresponding SE(3)-equivariant transformation $\rho(\boldsymbol{R}_i, \boldsymbol{t}_i)$ to the visual tokens $\boldsymbol{r}_{\text{2d}}$ by applying the induced SE(3)-equivariant action parameterized by $(\boldsymbol{R}_i, \boldsymbol{t}_i)$ to obtain the transformed representations $\boldsymbol{r}_{\text{se(3)}}$. The alignment module performs per-frame, per-token representation alignment between $\boldsymbol{r}_{\text{se(3)}}$ and $\boldsymbol{r}_{\text{3d}}$, and additionally enforces per-frame pose alignment between $\boldsymbol{r}_{\text{pose}}$ and $\hat{\boldsymbol{r}}_{\text{pose}}$. We specify the full geometry-aware alignment formulation in Section 3.2.

3.2 REPA FORMULATION WITH SE(3) STRUCTURE

SE(3) Notation and Equivariance. Our goal is to align token representations with 3D geometry under viewpoint changes. To make the alignment consistent across frames, we impose SE(3) equivariance so that applying a rigid motion in 3D induces a structured transformation in the representation space. Formally, let $g \in SE(3)$, and let $\rho_X(g)$ and $\rho_Y(g)$ denote the induced group actions on the input space X and output space Y. A mapping $f : X \to Y$ is SE(3)-equivariant if

$$f\big(\rho_X(g)\, x\big) \;=\; \rho_Y(g)\, f(x), \quad \forall g \in SE(3), \forall x \in X. \tag{2}$$

In our case, $\rho(\boldsymbol{R}_i, \boldsymbol{t}_i)$ specifies the induced SE(3) action applied to $\boldsymbol{r}_{\text{2d}}$. For simplicity, we omit the subscripts and write $\rho(\cdot)$ when the input and output actions are clear from context.

SE(3) Action on Visual Tokens. We predict per-frame SE(3) parameters from the pose tokens $\boldsymbol{r}_{\text{pose}}$ using two MLP heads, producing a rotation $\boldsymbol{R}_i \in \mathbb{R}^{3\times3}$ and a translation $\boldsymbol{t}_i \in \mathbb{R}^3$ for each frame. The rotation is parameterized in 6D and orthogonalized to obtain a valid rotation in SO3 (the 3D rotation group), which avoids the discontinuities and singularities associated with Euler angles.

To apply the induced SE(3) action to visual tokens, we first lift $\boldsymbol{r}_{\text{2d}}$ with a projection head into an intermediate representation $\tilde{\boldsymbol{r}} = \{\tilde{\boldsymbol{r}}_i\}_{i=1}^S$, where $\tilde{\boldsymbol{r}}_i \in \mathbb{R}^{N\times3K}$. The SE(3) transformation is

$$\rho(\boldsymbol{R}_i, \boldsymbol{t}_i)\, \tilde{\boldsymbol{r}}_i \;:=\; \tilde{\boldsymbol{r}}_i \boldsymbol{R}_i^\top + \mathbf{1}_{N\times K}\, \boldsymbol{t}_i^\top. \tag{3}$$

This operation is equivalent to a block-diagonal transform in the $3K$-dimensional projected space, with translation broadcast across all N tokens and K subspaces. We then map the transformed

features back and obtain

$$\boldsymbol{r}_{\text{se(3)}} = \{\,\text{MLP}(\rho(\boldsymbol{R}_i, \boldsymbol{t}_i)\,\tilde{\boldsymbol{r}}_i)\,\}_{i=1}^{S}. \tag{4}$$

which is used in subsequent geometry-aware representation alignment.

Geometry-Aware REPA. REPA advocates aligning intermediate hidden representations to strong external visual representations to improve representation learning. We adopt this principle in a geometry-aware manner by aligning MLLM hidden representations to geometry-centric targets produced by the Visual Geometry Encoder. Specifically, the alignment targets include both 3D geometry representations $\boldsymbol{r}_{\text{3d}}$ and pose representations $\hat{\boldsymbol{r}}_{\text{pose}}$. On the MLLM side, we align the SE(3)-aware visual representations $\boldsymbol{r}_{\text{se(3)}}$ (obtained by applying $\rho(\boldsymbol{R}_i, \boldsymbol{t}_i)$ to the projected visual tokens) with $\boldsymbol{r}_{\text{3d}}$, and align the pose tokens $\boldsymbol{r}_{\text{pose}}$ with $\hat{\boldsymbol{r}}_{\text{pose}}$. We optimize this geometry-aware alignment using a cosine similarity objective, with details provided in Section 3.3.

3.3 Learning Objectives

Geometry-Aware Alignment Objective. For visual representation alignment, we use the projection head h_ϕ to map SE(3)-aware tokens to the geometry embedding space, and define

$$\mathcal{L}_{\text{VIS}} = -\,\mathbb{E}_{\boldsymbol{I},Q}\,\mathbb{E}_{i,n}\Big[\,\text{sim}\big(h_\phi(\boldsymbol{r}_{\text{se(3)}}^{(i,n)}),\, \boldsymbol{r}_{\text{3d}}^{(i,n)}\big)\Big], \tag{5}$$

for pose-level alignment, we use another projection head h_φ and define

$$\mathcal{L}_{\text{POSE}} = -\,\mathbb{E}_{\boldsymbol{I},\,Q}\,\mathbb{E}_{i}\Big[\,\text{sim}\big(h_\varphi(\boldsymbol{r}_{\text{pose}}^{(i)}),\, \hat{\boldsymbol{r}}_{\text{pose}}^{(i)}\big)\Big], \tag{6}$$

where $i \in \{1, \ldots, S\}$ and $n \in \{1, \ldots, N\}$, S is the number of frames and N is the number of visual tokens per frame. We combine them as

$$\mathcal{L}_{\text{ALIGN}} = \mathcal{L}_{\text{VIS}} + \lambda_{\text{POSE}}\mathcal{L}_{\text{POSE}}. \tag{7}$$

Stage-wise Objectives. We optimize the model in two stages, supervised fine-tuning (SFT) and reinforcement learning (RL), with a simple cold start on generated chain-of-thought data to stabilize instruction following in the RL stage. Training details are provided in Section 4.1. In SFT and the cold start, we minimize the cross-entropy together with the geometry-aware alignment term $\mathcal{L}_{\text{ALIGN}}$, weighted by a coefficient λ_{ALIGN}. In the subsequent RL, we optimize $\pi_{\boldsymbol{\theta}}$ with GRPO (Shao et al., 2024; Guo et al., 2025) objective. The update ratio $r_{i,t}$ is the likelihood ratio between the current policy $\pi_{\boldsymbol{\theta}}$ and the old policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$ at token $o_{i,t}$, clipped with ε for stability. The advantage $\widehat{A}_{i,t}$ is computed using relative rewards within each group. The policy loss is

$$\mathcal{L}_{\pi} = \mathbb{E}_{i,t}\Big[\min\Big(r_{i,t}\,\widehat{A}_{i,t},\, \text{clip}\big(r_{i,t},\, 1 \pm \varepsilon\big)\,\widehat{A}_{i,t}\Big)\Big], \tag{8}$$

where $t = 1 : T_i$ and $i = 1 : G$, with G denoting the group size, the update ratio

$$r_{i,t} = \frac{\pi_{\boldsymbol{\theta}}\big(o_{i,t} \,\big|\, \boldsymbol{o}_{i,1:t-1},\, \boldsymbol{I},\, Q\big)}{\pi_{\boldsymbol{\theta}_{\text{old}}}\big(o_{i,t} \,\big|\, \boldsymbol{o}_{i,1:t-1},\, \boldsymbol{I},\, Q\big)}, \tag{9}$$

and the token $o_{i,t} = y_t^{(i)}$. GRPO further regularizes with a KL divergence term:

$$\mathcal{L}_{\text{GRPO}} = \mathcal{L}_{\pi} - \beta\,\mathbb{D}_{\text{KL}}[\,\pi_{\boldsymbol{\theta}}\,||\,\pi_{\text{ref}}\,], \tag{10}$$

where β is the coefficient of the KL penalty.

4 Experiments

4.1 Experimental Setup

Backbones and Training Strategy. We pair Qwen2.5-VL-3B/7B with VGGT-1B as a training-time teacher; Qwen3-VL-8B retains the teacher, objectives, and training strategy on Scan2Cap. A 3B Scan2Cap ablation also replaces VGGT with DA3. Following VG-LLM, supervised fine-tuning (SFT) uses 234K SPAR-7M geometry and 63K LLaVA-Hound video samples. Cold start uses 3DThinker-supplied GPT-4o chain-of-thought (CoT) data, cleaned, filtered, and VGGT-index-aligned, to stabilize instruction following during reinforcement learning (RL) with GRPO. We found

Table 1: **Evaluation on VSI-Bench.** GEO-REPA shows highly competitive performance compared to existing methods, demonstrating the effectiveness of the proposed geometry-aware alignment.

Methods	Year	Numerical Question				Multiple-Choice Question				Average	Rank
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order		
Proprietary Models (API)											
Gemini-3-Pro	2025	49.0	42.8	71.5	41.8	56.6	57.5	61.9	60.0	55.1	–
GPT-5	2025	53.3	34.4	73.3	47.5	63.7	48.6	50.2	68.9	55.0	–
Kimi-K2.5	2026	57.2	34.9	69.3	54.4	59.6	41.3	52.1	67.0	54.5	–
Open-Source Models											
InternVL2-40B	2024c	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6	36.0	5
LLaVA-OneVision-72B	2025a	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2	3
LongVILA-8B	2025	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5	21.6	6
BAGEL-7B	2025a	62.8	36.3	56.4	49.7	46.1	49.4	26.8	43.1	46.3	1
LLaVA-Video-72B	2025b	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9	2
Qwen2.5-VL-72B	2025	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9	37.0	4
Spatial-Enhanced Models											
SAT-LLaVA-Video-7B	2024	–	–	–	47.3	41.1	37.1	36.1	40.4	–	–
SPAR-8B	2025a	–	–	–	–	–	–	–	–	41.1	17
Spatial-MLLM-4B	2025a	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3	48.4	12
VG-LLM-4B	2025a	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4	47.3	14
VG-LLM-8B	2025a	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2	50.7	11
3DRS-7B	2025b	68.7	34.8	53.6	56.6	40.9	43.2	30.4	39.2	45.9	15
3DThinker-4B	2026b	–	–	–	–	–	–	–	–	59.1	8
3DThinker-8B	2026b	–	–	–	–	–	–	–	–	63.7	3
VLM-3R-7B	2026	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1	60.9	6
SpaceR-7B	2025	–	–	–	–	–	–	–	–	45.6	16
GS-Reasoner-8B	2026a	69.1	61.9	70.0	65.7	65.4	88.9	44.3	52.3	64.7	2
VST-3B	2026a	66.6	45.0	72.8	60.9	59.9	47.6	40.7	68.3	57.7	9
VST-7B	2026a	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2	61.2	5
Stream3D-VLM-4B	2026	68.6	38.4	65.9	56.4	49.9	54.3	42.3	66.0	55.2	10
VLM-IE3D-4B	2026	67.5	38.5	55.0	58.7	47.7	45.1	36.6	31.9	47.6	13
FactoSR-8B	2026b	66.4	46.7	75.7	63.7	60.3	–	46.9	65.7	61.5	4
GEO-REPA-3B (Ours)	2026	66.4	51.2	66.2	61.3	57.9	77.1	41.3	53.6	59.4	7
GEO-REPA-7B (Ours)	2026	70.5	58.3	70.4	67.1	66.2	85.1	47.6	58.0	65.4	1

no shared image filenames between cold-start and evaluation data. Alignment is trained in SFT/cold start; benchmark-specific GRPO updates only the LLM on each task's training split, without alignment loss. VSI-Bench/SQA3D rewards follow Spatial-MLLM; Scan2Cap uses CIDEr and format rewards. Alignment ablations match data, prompts, frame sampling, rewards, and optimization across stages. Inference removes VGGT, the SE(3) predictor, pose tokens, and alignment heads.

Datasets and Metrics. VSI-Bench uses multiple-choice accuracy and numerical Mean Relative Accuracy; SQA3D uses Exact Match (EM-1/EM-R1); Scan2Cap uses CIDEr, BLEU-4 (B-4), METEOR, and ROUGE. We also test general ability on MMMU, TextVQA, MMBench, and MM-Vet, and pose accuracy against ScanNet validation extrinsics.

Implementation Details. Training uses 8 NVIDIA A100-80GB GPUs and AdamW (Loshchilov & Hutter, 2019) (weight decay 0.01, 3% warmup) across stages. We sample 16 frames randomly for training and uniformly for inference; aspect-preserving image resizing yields 32–192 visual tokens, and video frames match the first frame's dimensions. For Qwen2.5-VL, we match layer-8

Table 2: **Evaluation on Scan2Cap.** GEO-REPA-7B demonstrates superior dense captioning capabilities, outperforming prior work on most metrics.

Methods	Year	CIDEr	B-4	Meteor	Rouge
Scan2Cap	2021	39.1	23.3	22.0	44.8
LL3DA	2024a	65.2	36.8	–	–
Chat-Scene	2024	77.1	36.3	28.0	58.1
LLaVA-3D	2024	84.1	42.6	29.0	63.4
Grounded 3D-LLM	2024b	70.6	35.5	–	–
Video-3D LLM	2025b	83.8	41.3	–	–
VG-LLM-4B	2025a	78.6	40.9	28.6	62.4
VG-LLM-8B	2025a	80.0	41.5	28.9	62.6
Ross3D-7B	2025a	81.3	43.4	30.3	66.9
3DRS-7B	2025b	86.1	41.6	–	–
GS-Reasoner-8B	2026a	101.0	47.6	32.1	69.2
GEO-REPA-3B	2026	80.1	42.3	32.1	63.6
GEO-REPA-7B	2026	90.2	48.0	33.5	69.8

Table 3: **Effect of representation alignment and SE(3)-equivariant representations.** Green denotes our method.

Methods	CIDEr	B-4	Meteor	Rouge
Qwen2.5-VL-3B				
Baseline	61.7	30.2	25.9	51.8
+ Naive Alignment	72.3	37.3	28.1	57.7
Δ *Improvement*	*(+10.6)*	*(+7.1)*	*(+2.2)*	*(+5.9)*
+ SE(3) Alignment	80.1	42.3	32.1	63.6
Δ *Improvement*	*(+18.4)*	*(+12.1)*	*(+6.2)*	*(+11.8)*
Qwen2.5-VL-7B				
Baseline	73.2	35.8	27.8	58.0
+ Naive Alignment	82.3	42.5	29.3	63.0
Δ *Improvement*	*(+9.1)*	*(+6.7)*	*(+1.5)*	*(+5.0)*
+ SE(3) Alignment	90.2	48.0	33.5	69.8
Δ *Improvement*	*(+17.0)*	*(+12.2)*	*(+5.7)*	*(+11.8)*

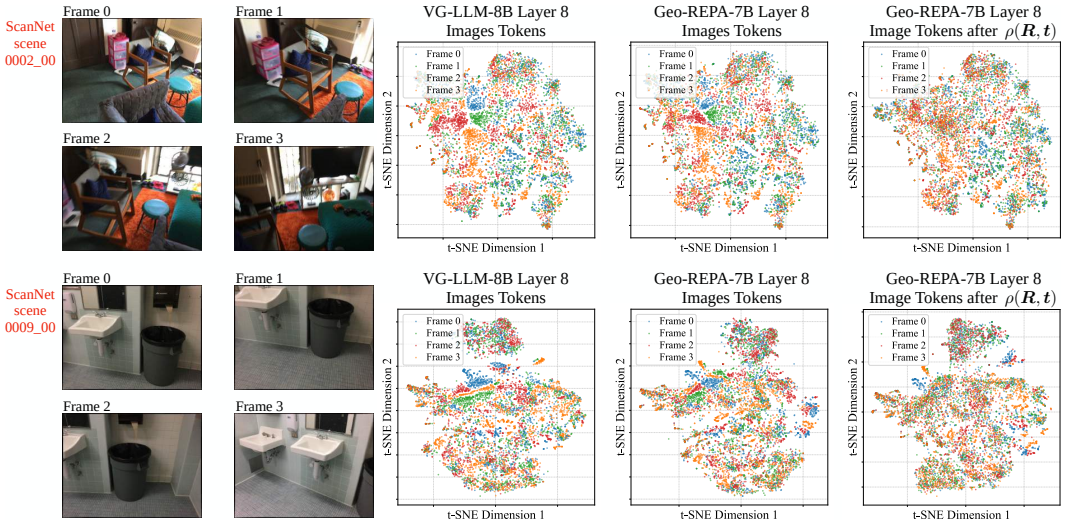


Figure 4: **Visualization of SE(3)-aware representations.** We visualize layer-8 image tokens from two ScanNet scenes, where different colors denote frames under different camera poses. After applying the predicted SE(3) transformation $\rho(\mathbf{R}, \mathbf{t})$, tokens from different frames become more mixed and closer in the representation space, indicating reduced pose-dependent separation.

features with VGGT's final tokens in dimension and resolution; geometry and pose representations share frame 0 and sequence scale. Token matching, normalization, prompts, decoding, and answer parsing follow Spatial-MLLM. LLM learning rates are 8×10^{-6} (constant, SFT/cold start) and 2×10^{-6} (cosine, GRPO); alignment-module rates are $3\times/1\times$ the LLM rate in SFT/cold start. We set $\lambda_{\text{ALIGN}} = 0.6$, $\lambda_{\text{POSE}} = 1.0$, and default $G = 8$; see the group-size ablation in Figure 6a. Separate matched latency tests use eight H800 GPUs (BF16), one scene/GPU, 16 uniform frames, 196 visual tokens/frame, and identical prompts, decoding, and generation limits.

4.2 Overall Performance

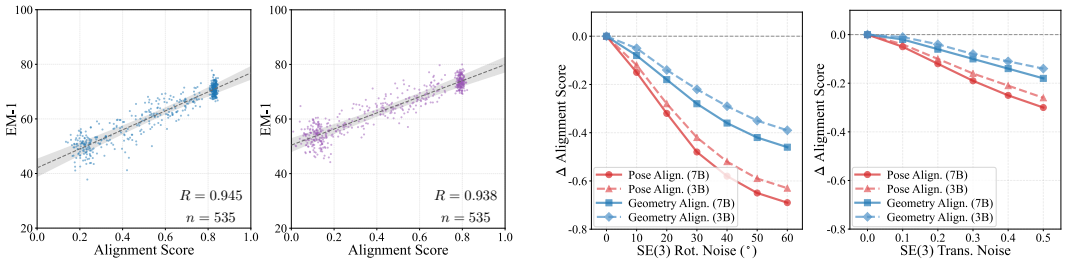
Results on VSI-Bench. As shown in Table 1, Geo-REPA-7B achieves state-of-the-art performance with an average score of 65.4, surpassing both general MLLMs and spatial specialists such as GS-Reasoner-8B and 3DThinker-8B. The gains on geometry tasks, including *Object Size* (70.4), *Room Size* (67.1), and *Relative Distance* (66.2), show the benefit of geometry-aware alignment for spatial reasoning. Notably, Geo-REPA-3B also outperforms much larger models such as LLaVA-Video-72B and Qwen2.5-VL-72B, suggesting an efficient path to improving spatial capability.

Results on SQA3D. Table 4 evaluates egocentric 3D scene understanding and reasoning. Geo-REPA-7B leads video-input models with 63.6 EM-1. Using only RGB frames, it outperforms explicit 3D-input methods, including LLaVA-3D and Video-3D LLM, showing that our alignment can improve spatial reasoning without depth inputs. The lightweight Geo-REPA-3B substantially outperforms Qwen2.5-VL-72B, showing the effectiveness of targeted geometry-aware alignment.

Results on Scan2Cap. Dense captioning requires 3D object localization and attribute description. In Table 2, Geo-REPA-7B achieves best performance on BLEU-4, METEOR, and ROUGE, outperforming explicit 3D-input models including Video-3D LLM. Despite GS-Reasoner's higher CIDEr, our stronger linguistic metrics suggest improved attribute description and spatial consistency. Geo-REPA-3B matches the second-best METEOR score and outperforms larger VG-LLM-8B.

Table 4: **Evaluation on SQA3D.** Geo-REPA achieves competitive performance against baselines.

Methods	Year	EM-1	EM-R1
3D/2.5D-Input Models			
LLaVA-3D	2024	60.1	–
Chat-Scene	2024	54.6	57.5
3D-LLaVA	2025b	54.5	56.6
Video-3D LLM	2025b	58.6	–
Ross3D-7B	2025a	63.0	65.7
Video-Input Models			
Qwen2.5-VL-72B	2025	47.0	50.9
SPAR-7B	2025a	58.1	–
BAGEL-7B-FT	2025a	57.2	59.7
LLaVA-Video-7B	2025b	48.5	–
Spatial-MLLM-4B	2025a	55.9	58.7
3DRS-7B	2025b	60.6	–
GS-Reasoner-8B	2026a	59.9	–
Geo-REPA-3B	2026	56.8	60.2
Geo-REPA-7B	2026	63.6	65.4



(a) Correlation analysis. (b) SE(3) perturbation analysis.

Figure 5: **Correlation and SE(3) perturbation analysis.** (a) On held-out SQA3D validation batches, we plot EM-1 against visual-geometry (Left) and pose (Right) alignment scores. Pearson correlations ($R > 0.93$) indicate a strong association. (b) We inject rotation and translation noise into predicted poses. Decreasing alignment indicates sensitivity to the supplied transformations.

Table 5: **Qwen3-VL-8B on Scan2Cap.** CIDEr scores for the backbone and Geo-REPA variant.

Backbone	Variant	CIDEr ↑
Qwen3-VL-8B	Baseline	79.7
Qwen3-VL-8B	Geo-REPA	**94.1**

Table 6: **General Multimodal Performance.** Comparison with Qwen2-VL-7B and the Qwen2.5-VL-7B backbone under identical evaluation settings.

Model	MMMU ↑	TextVQA ↑	MMBench ↑	MMVet ↑
Qwen2-VL-7B	54.1	84.3	80.7	62.0
Qwen2.5-VL-7B	58.6	84.9	82.6	67.1
Geo-REPA-7B	57.3	84.2	79.8	64.5

4.3 Analysis of Geometry-Aware Alignment

Visualization of Representations. Using the final checkpoint, we visualize layer-8 image tokens from ScanNet scenes under different camera poses. As shown in Figure 4, tokens from different frames are more separated in VG-LLM, whereas Geo-REPA produces a more organized distribution. Applying the predicted transformation $\rho(\mathbf{R}, \mathbf{t})$ further mixes tokens across frames, consistent with reduced pose-dependent separation in the learned alignment space.

Correlation Analysis of Alignment. We analyze how visual-geometry and pose alignment scores relate to SQA3D EM-1 on held-out validation batches from frozen checkpoints (Figure 5a). The ordinary least squares fit is displayed with an illustrative band $\hat{y}(x) \pm 10\,\mathrm{s.e.}(\hat{y}(x))$. The factor 10 affects neither the fit nor Pearson correlation, and the band is not a calibrated confidence interval. Both alignment scores show strong positive correlations ($R > 0.93$), supporting an association with spatial performance without establishing causality.

SE(3) Perturbation Sensitivity. We rotate predicted poses about a random unit axis by $\delta\theta$ or add a displacement of relative magnitude $\|\delta\mathbf{t}\|/\|\mathbf{t}\|$. In Figure 5b, alignment decreases as pose noise increases. Pose alignment drops more sharply than visual-geometry alignment, and the 7B model shows greater sensitivity than the 3B model in this test. These declines show sensitivity to supplied transformations. We evaluate the predicted camera motion against ground truth below.

Predicted Camera Pose Accuracy. Using the frozen final checkpoint, we compare predicted transformations with ScanNet validation extrinsics relative to frame 0, without probe or scale fitting (Table 8). Against shuffled poses, rotation error falls from $47.8°$ to $8.6°$ and translation-direction error from $78.4°$ to $14.7°$. Translation-magnitude error decreases from 59.1% to 27.0%. The remaining magnitude error limits claims about accurate scale, although the predictions capture substantial orientation and direction information.

Backbone Generalization and General Multimodal Performance. On Scan2Cap, applying Geo-REPA to Qwen3-VL-8B raises CIDEr from 79.7 to 94.1 (Table 5), showing that the benefit extends to this newer backbone. On MMMU, TextVQA, MMBench, and MMVet, Geo-REPA-7B trails its Qwen2.5-VL-7B backbone by 1.3, 0.7, 2.8, and 2.6 points, respectively (Table 6). These consistent decreases indicate a limited general-performance trade-off after spatial post-training. The comparison does not isolate the effects of alignment and the spatial SFT mixture.

Table 7: **Matched Alignment Ablation on VSI-Bench and SQA3D.** Geo-REPA-3B results with identical training settings.

Design	VSI-Bench Avg.	SQA3D EM-1	SQA3D EM-R1
No alignment	52.6	51.5	55.0
Naive alignment	56.7	54.3	57.9
SE(3) alignment	**59.4**	**56.8**	**60.2**

Table 8: **Camera Pose Errors on ScanNet.** Frozen-model errors against ground-truth extrinsics relative to frame 0.

Error metric ↓	Shuffled (R, t)	Geo-REPA (R, t)
Rotation geodesic (°)	47.8	**8.6**
Translation direction (°)	78.4	**14.7**
Translation magnitude (%)	59.1	**27.0**

4.4 Ablation Study

Effect of SE(3) Alignment. Within each benchmark, we compare no alignment, naive direct visual-to-geometry alignment, and pose-aware SE(3)-structured alignment with identical SFT and cold-start data, frame sampling, rewards, and task-specific GRPO settings. On Scan2Cap, the full design outperforms both alternatives (Table 3). The matched 3B experiments show the same ordering across no, naive, and SE(3) alignment: VSI-Bench Avg. rises from 52.6 to 56.7 to 59.4, and SQA3D EM-1 from 51.5 to 54.3 to 56.8 (Table 7). Results support the pose-aware alignment design within its auxiliary branch, without establishing exact SE(3) equivariance of the entire MLLM.

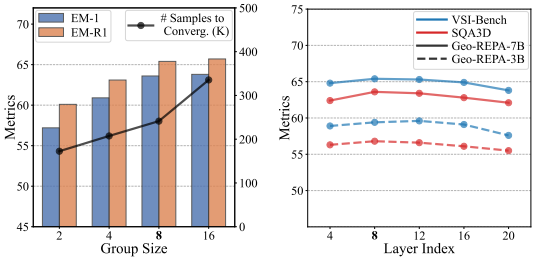
Training Stages. For the full Geo-REPA-7B model on Scan2Cap, cold start raises CIDEr from 82.2 to 84.0, and GRPO further raises it to 90.2 (Table 9 reports the endpoints). The no- and naive-alignment variants also benefit from GRPO on this task. Separately, on held-out samples, alignment decreases from 0.85 after SFT to 0.58 after GRPO but remains above the no-alignment score of 0.07. Task performance thus improves even as GRPO partially weakens the measured geometric alignment.

GRPO Group Size. Figure 6a compares group size G by downstream performance and samples to convergence. Gains diminish as G increases, while sampling cost per update grows. We use $G = 8$ as a practical trade-off.

Alignment Layer Index. In Figure 6b, intermediate-to-early layers perform best among the tested indices, consistent with REPA (Yu et al., 2024), while nearby choices remain competitive in both models. We use layer 8 for the evaluated Qwen2.5-VL backbones without assuming that the same index is optimal at untested larger model depths.

Table 9: **Effect of Training Stages.** Green denotes our default settings on Scan2Cap.

Training Stages	CIDEr	B-4	Meteor	Rouge
Geo-REPA-3B				
SFT Only	78.3	40.2	28.7	61.5
SFT + RL	80.1	42.3	32.1	63.6
Δ *Improvement*	*(+1.8)*	*(+2.1)*	*(+3.4)*	*(+2.1)*
Geo-REPA-7B				
SFT Only	82.2	42.5	29.7	63.1
SFT + RL	90.2	48.0	33.5	69.8
Δ *Improvement*	*(+8.0)*	*(+5.5)*	*(+3.8)*	*(+6.7)*



(a) Effect of GRPO Group Size. (b) REPA Layer Sensitivity.

Figure 6: **Ablation study** of performance and efficiency trends under varying parameter settings.

5 Conclusion

In this work, we introduced Geo-REPA to help MLLMs retain and use 3D cues beyond what input-level fusion alone provides. It uses predicted SE(3) transformations to account for camera motion, aligning intermediate visual tokens with 3D features and pose tokens with camera pose targets during training. Across VSI-Bench, SQA3D, and Scan2Cap, Geo-REPA achieves competitive or leading results, and its Scan2Cap gain extends to a newer MLLM backbone. Frozen-checkpoint diagnostics, cross-view visualizations, and quantitative alignment and pose analyses suggest retained 3D cues and sensitivity to camera motion. Matched ablations across all three tasks show that pose-aware SE(3) alignment improves performance over no alignment and direct feature matching. The trained model needs no external 3D encoder at inference, although general multimodal scores decline modestly relative to the underlying backbone. Together, these findings support geometry-aware supervision for spatial reasoning and motivate further work on preserving broader capabilities.

LLM USAGE STATEMENT.

LLMs were utilized as auxiliary tools for organizing code, checking the consistency of experimental comparisons, and improving the clarity of the manuscript. The research ideas, experimental design, reported results, and corresponding analyses were independently conducted and verified by the authors, who are fully responsible for the content of this work.

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.

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*, pp. 26428–26438, 2024a.

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

Yiming Chen, Zekun Qi, Wenyao Zhang, Xin Jin, Li Zhang, and Peidong Liu. Reasoning in space via grounding in the world. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. LongVILA: Scaling long-context visual language models for long videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

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, 2026b.

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

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3193–3203, 2021.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025a.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3772–3782, 2025b.

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.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3DRS: MLLMs need 3d-aware representation supervision for scene understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

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*, 2025b.

Haoyu Li, Mingyang Han, Yu Xi, Dongxiao Wang, Hankun Wang, Haoxiang Shi, Boyu Li, Jun Song, Bo Zheng, and Shuai Wang. Time-layer adaptive alignment for speaker similarity in flow-matching based zero-shot tts. *arXiv preprint arXiv:2511.09995*, 2025c.

Wenhao Li, Xueying Jiang, Quanhao Qian, Deli Zhao, Ran Xu, Shijian Lu, and Gongjie Zhang. 3d-aware vlms with implicit and explicit geometries. In *European Conference on Computer Vision*, pp. 168–185. Springer, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Nicolas Sereyjol-Garros, Ellington Kirby, Victor Besnier, and Nermin Samet. Leveraging 3d representation alignment and rgb pretrained priors for lidar scene generation. *arXiv preprint arXiv:2601.07692*, 2026.

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.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

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

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

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

Lintong Wei, Jian Lu, Haozhe Cheng, Jihua Zhu, and Kaibing Zhang. Point-sra: Self-representation alignment for 3d representation learning. *arXiv preprint arXiv:2601.01746*, 2026.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025b.

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*, pp. 10632–10643, 2025a.

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*, pp. 10632–10643, 2025b.

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. In *European Conference on Computer Vision*, pp. 192–211. Springer, 2026a.

Yijun Yang, Shenghe Zheng, Wenbo Li, Jianhui Liu, Haoze Sun, Yanbing Zhang, Jiaxiu Jiang, Lin Song, Haoyang Huang, Nan Duan, et al. Unfold the world: Factorize 4d properties in reinforcing spatial reasoning. In *European Conference on Computer Vision*, pp. 605–622. Springer, 2026b.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Hanxun Yu, Xuan Qu, Lei Ke, Boqiang Zhang, Yuxin Wang, Jianke Zhu, and Dong Yu. Stream3d-vlm: Online 3d spatial understanding with incremental geometry priors. In *European Conference on Computer Vision*, pp. 130–148. Springer, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

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*, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing MLLMs with 3d vision geometry priors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

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*, pp. 8995–9006, 2025b.

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.