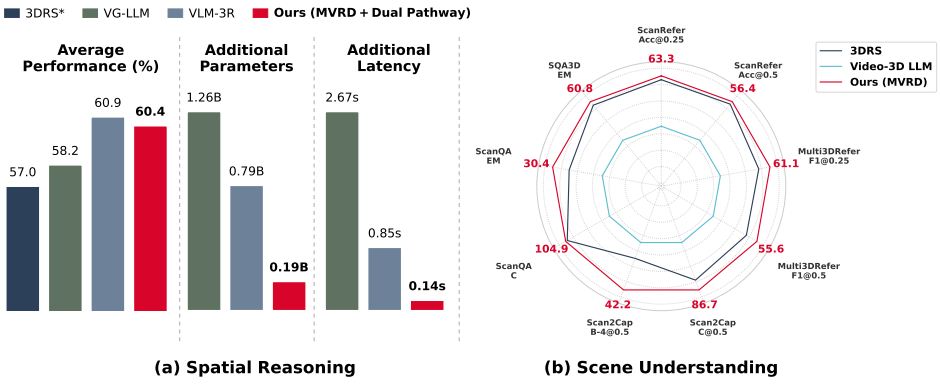
Multi-View Relational Distillation for Spatial Reasoning with Vision-Language Models

Kiet T. Nguyen **Hanbo Shim** **Jinwoo Kim**[†] **Seunghoon Hong**[†]
KAIST

Abstract

Vision-language models (VLMs) have achieved strong image and video understanding, yet their visual-spatial representations remain geometrically fragile, leading to failures in spatial reasoning needed for embodied AI, robotics, and autonomous driving. Prior approaches to geometry grounding either fine-tune VLMs on spatial question answering, which can perpetuate spurious visual representations, or fuse features from large geometry-grounded vision models, which substantially increases model size at inference. Knowledge distillation from geometry-grounded vision models offers an alternative, but directly matching multi-view teacher features can disrupt the pretrained alignment between visual and textual representations, degrading object- and language-semantic capabilities. We propose **multi-view relational distillation (MVRD)**, which distills patch-wise cosine similarities across views instead of the teacher features themselves. These relations encode geometric correspondences adequate for spatial understanding, while leaving the student representation underdetermined, allowing it to remain close to its pretrained vision-language space. Across representative VLMs, MVRD improves visual-spatial reasoning, outperforming supervised fine-tuning and feature distillation while approaching feature fusion methods with considerably less added parameters and lower latency. We show that MVRD makes visual representations more geometric while retaining language alignment, and generalizes to 3D scene understanding tasks such as object grounding, dense captioning, and question answering. Code is available at `https://github.com/kietngt00/MVRD`



(a) Spatial Reasoning **(b) Scene Understanding**

Figure 1: **Multi-view relational distillation (MVRD). (a) Spatial reasoning.** MVRD with Dual Pathway variant approaches the feature fusion method VLM-3R [1], using the same backbone and dataset, with far fewer added parameters and lower latency. **(b) Scene understanding.** MVRD transfers to scene understanding, surpassing the prior state-of-the-art 3DRS [2] in all tasks.

[†]Equal advising

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

1 Introduction

Humans perceive, understand, and act upon physical environments through rich visual-spatial representations, paired with language as a flexible abstraction for describing scenes, specifying goals, and composing intentions. Evidence from cognitive science and neuroscience suggests that perception and action rely on structured, manipulable representations of visual scenes [3–5], supporting capabilities such as object localization [6], mental viewpoint changes [7], and path planning [8]. This complementarity motivates vision-language models (VLMs) that not only align images with text, but also learn geometrically faithful visual representations for grounded spatial reasoning and action, with applications ranging from embodied AI and robotics [9–12] to autonomous driving [13].

Although current VLMs are trained with profound scales of data and compute [14–16], it is known that their visual-spatial representations are fractured [17], leading to failures in geometric tasks as simple as inferring relative distances [18–20]. As such, there has been substantial effort in grounding VLMs to geometry. Early efforts perform supervised fine-tuning (SFT) with spatial question answering [21–23], which do not directly address the problem of spurious representations [24]. More recent work propose to fuse multi-view features from a large geometry-grounded vision model [25, 26, 1, 27], however at the cost of substantially increasing the already large model size during inference time.

Here, we aim to improve spatial representation and understanding in VLMs while minimally increasing their size. Prior work has proposed knowledge distillation [28] as a solution, matching multi-view features of VLMs to be similar with those of a geometry-grounded vision model [2]. We identify a problem of this idea: while it grounds VLMs' visual representations, it obstructs their pretrained alignment with language. We find that this degrades capabilities of linguistic grounding of objects, e.g. tracking the order in which objects of a given class appear. Prior work across multiple domains [29–32] has also shown that distillation from off-the-shelf teachers can disrupt a student model's pretrained representations. This motivates our key question: *How can we distill geometric knowledge into VLMs without trading off vision-language alignment?*

To this end, we propose **multi-view relational distillation**, a simple method for improving spatial reasoning in VLMs via knowledge distillation from geometry-grounded vision models. Instead of distilling multi-view features themselves, we distill *cosine similarities* of features across views. Our hypothesis is that multi-view relations encode geometric correspondences adequate to improve spatial understanding of VLMs, yet allow visual representations to remain close to the pretrained vision-language space. The former is supported by evidence of relational processing in human vision [33–36]. The latter follows as multi-view similarities do not uniquely determine student features, so there can exist ones with low loss that lie close to the pretraining regime. Our contributions are as follows:

(1) **Failures of feature distillation.** Through evaluations and representation probing, we uncover a previously undiagnosed failure mode of feature distillation, resulting in degradations in vision-language alignment and related tasks despite making VLMs more geometrically grounded.

(2) **Multi-view relational distillation (MVRD).** Motivated by adequacy and minimality aspects of relational visual representations, we propose MVRD, an approach that distills similarities within multi-view features from geometry-grounded vision models into VLMs for geometry grounding.

(3) **Competitive performance.** MVRD outperforms SFT and feature distillation at least by 2.2 points in accuracy on VSI-Bench, and its best setting narrows the gap to the feature fusion counterpart to 0.5 points. We show that MVRD meaningfully retains vision-language alignment.

(4) **Computational efficiency.** Being a distillation method, MVRD introduces less than 25% additional parameters and less than 16% runtime overhead compared to its feature fusion counterpart.

2 Related work

VLMs and their fractured spatial understanding. VLMs that process text jointly with multiple images, such as multiple views of a scene or video frames, have achieved strong performances in classical benchmarks [14–16, 37, 38]. However, recent findings, including from VSI-Bench [19], have shown that these models are fragile in multi-view spatial reasoning, which requires inferring geometric structures such as egocentric-allocentric transformations [39, 40]. This can be attributed to their image encoders [41, 42], which produce representations with limited 3D-awareness needed for depth estimation or matching multi-view correspondences [17]. The current paradigm, which

fine-tunes VLMs with visual question answering, has been reported to fall short of resolving these spurious visual representations [43, 44].

Grounding VLMs to visual geometry. There have been considerable efforts in grounding VLMs to spatial geometry. A line of work proposed to perform SFT on spatial question-answer pairs [45–47]. However, SFT does not directly address the underlying problem of fractured visual representations, and as such, we show in our setting that it alone leads to limited gains. Recent work tackles this problem at the representation level, leveraging recent large-scale geometry-grounded vision foundation models such as VGGT [48] and CUT3R [49]. *Feature fusion* methods such as VLM-3R [1], VG-LLM [26], and SpatialStack [50] take their features and inject them to the visual pathway of VLMs, with G^{2}VLM [51] further takes a dual-expert approach, duplicating the model and performing feature fusion in one of them. While these methods demonstrate the benefit of geometry-grounded representations in VLMs, they require a standalone model for feature extraction at inference time, substantially increasing net model size, posing a challenge in embodied applications.

Feature distillation methods. Given the above, knowledge distillation [28] methods that operate at the representation level offer an alternative route towards geometry-grounded VLMs while minimally increasing their size. Importantly, 3DRS [2] directly aligns multi-view visual features of VLMs with those from the geometry-grounded vision model VGGT. Relatedly, 3DThinker [52] and VaLR [53] align latent chain-of-thought features using curated data and multi-stage training. While 3DRS has been shown to improve geometry grounding of VLMs, and others have gained improvement via chain-of-thought training, how such geometric alignment affects the pretrained language-aligned visual space remains underexplored. Through careful evaluations and analysis, we find that feature distillation disrupts the alignment between visual and linguistic representations in VLMs, causing degradations in the tasks of estimating the object size or tracking the appearance order of objects.

Relational representations and distillation. Relational representations have been demonstrated to be useful in various contexts [54–56]. In human vision, evidence suggest that relational information across temporal stream of visual inputs is leveraged in spatial understanding and object perception [33–36]. In deep learning, a line of work has studied relational knowledge distillation [56], which has been shown to be effective in the vision domain [57, 31]. While these prior work consider relations between images or patches in an image, we extend these approaches and consider relations within multiple views of a scene. Furthermore, we argue that relational distillation better sustains vision-language alignment in VLMs based on the fact that different underlying features can lead to the same relational representation, a property that has not been explicitly leveraged in the prior work.

3 Failures of geometric feature distillation for VLMs

3.1 Problem setup

VLMs in multi-view setting. For multi-view scene and video understanding, a VLM takes as input a collection of images $\mathbf{x}_1, ..., \mathbf{x}_T \in \mathbb{R}^{H\times W\times 3}$ with a language query $\mathbf{q}$. A vision encoder maps each image individually into patch-level features, $\mathbf{x}_t \mapsto \mathbf{v}_t \in \mathbb{R}^{H_p\times W_p\times d}$, which we refer to as vision tokens. $\mathbf{q}$ is encoded using subword embeddings from a language model, $\mathbf{q} \mapsto \mathbf{w} \in \mathbb{R}^{N\times d}$ with length N, which we refer to language tokens. The tokens are concatenated into $\mathbf{z} = (\mathbf{v}_1, ..., \mathbf{v}_T, \mathbf{w}) \in \mathbb{R}^{(TH_pW_p+N)\times d}$ and processed by the language model for downstream tasks. At l-th layer, we denote the features of these tokens as $\mathbf{z}^l = (\mathbf{v}_1^l, ..., \mathbf{v}_T^l, \mathbf{w}^l)$, where $\mathbf{z}^0 = \mathbf{z}$. Since VLMs tend to focus on aligned use of vision and language, they are often fragile on geometric and spatial tasks. To overcome this, early works performed SFT with spatial question answering, minimizing the loss $\mathcal{L}_{\mathsf{SFT}} \coloneqq -\log p(\mathbf{a} \mid \mathbf{x}_{1:T}, \mathbf{q})$ where p denotes the likelihood of gold answer $\mathbf{a}$ evaluated by the VLM.

Geometry-grounded features. Recent work has proposed to improve spatial understanding in VLMs with representations from a geometry-grounded vision model. Such a model jointly encodes multi-view images into features, $\mathbf{x}_{1:T} \mapsto \mathbf{g}_{1:T} \in \mathbb{R}^{T\times H_g\times W_g\times d_g}$, and is trained with specialized objectives such as point cloud and camera pose estimation. As a result, their multi-view features faithfully encode the underlying 3D geometry. Feature fusion methods directly incorporate these into the forward pathway of VLMs, either via resampling $\mathbf{g}_t \mapsto \bar{\mathbf{g}}_t$ into $H_p \times W_p$ grid followed by additions into vision tokens, or through cross-attention. These methods directly address the problem of spurious representations, and yield substantial gains on spatial tasks when combined with SFT. However, they require a standalone vision model for feature extraction, increasing model size.

Table 1: **Feature distillation vs. SFT** on *LLaVA-Video-7B-Qwen2* [16] using VLM-3R [1] dataset, evaluated on VSI-Bench. Originally reported results are included in Appendix E.1. * indicates training on the VLM-3R dataset.

Methods	Avg.	*Abs. Dist.*	*Room Size*	*Rel. Dist.*	*Rel. Dir.*	*Route Plan*	*Obj. Count*	*Obj. Size*	*Appr. Order*
		Geometric					**Linguistic grounding**		
SFT [16]	**57.7**	43.6	63.7	64.9	68.9	**40.7**	**70.6**	**70.8**	**38.5**
SFT + feature distillation (3DRS*) [2]	57.0	**51.3**	**69.1**	**65.8**	**74.5**	37.6	70.0	68.0	19.6

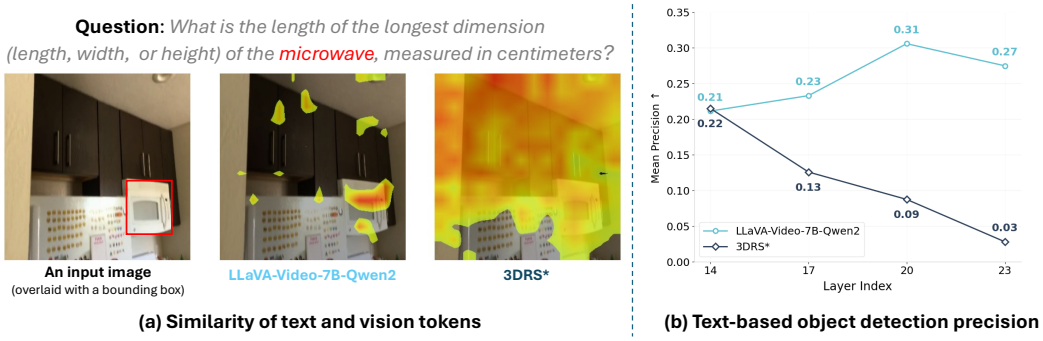


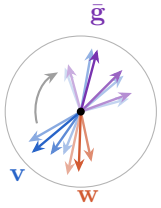
Figure 2: **Failure of feature distillation.** **(a)** *Qualitative*: At layer 20, 3DRS* fails to align *microwave* with the corresponding visual patches, unlike the base model LLaVA-Video-7B-Qwen2. **(b)** *Quantitative*: A sharp drop in text-based object-detection precision across layers indicates severe degradation in vision-language alignment.

Feature distillation. Feature distillation methods train VLMs to match the (resampled) geometric features $\bar{\mathbf{g}}_{1:T}$, then discard the additional vision model at inference time, thereby avoiding the overhead of feature fusion. This is done by minimizing the following loss for each input $(\mathbf{x}_{1:T}, \mathbf{q})$:

$$\mathcal{L}_{\mathsf{FD}} \coloneqq \frac{1}{TH_pW_p} \sum_{t=1}^{T} \sum_{h=1}^{H_p} \sum_{w=1}^{W_p} \ell_{\mathsf{FD}}\left(f\left(\mathbf{v}_t^l[h,w]\right), \mathsf{sg}(\bar{\mathbf{g}}_t[h,w])\right), \tag{1}$$

where $\ell_{\mathsf{FD}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a choice of error function, such as cosine distance, $f : \mathbb{R}^d \to \mathbb{R}^{d_g}$ is a position-wise MLP projection, and $\mathsf{sg}(\cdot)$ is the stop-gradient operator. The VLM and the projector are trained together with a composite loss $\mathcal{L}_{\mathsf{SFT}} + \lambda\mathcal{L}_{\mathsf{FD}}$ with hyperparameter $\lambda > 0$. The composite loss maximizes the likelihood of correct question answering by the VLM, while attempting to make its representations more geometric, by maximizing similarity, aiming for $\mathbf{v}_t^l \approx \bar{\mathbf{g}}_t$.

3.2 Failures of feature distillation for geometric grounding



The feature distillation loss in (1) usually enjoys stable optimization properties, and hence achieves the desired grounding $\mathbf{v}_t^l \approx \bar{\mathbf{g}}_t$ with ease. However, we identify its potential problem, obstructing the pretrained alignment between vision and language tokens. We illustrate this problem using a toy example in the figure on the right, showing three sets of feature vectors ($\mathbf{v}$, $\mathbf{w}$, $\bar{\mathbf{g}}$) in the two-dimensional plane. As feature distillation (arrow) brings vision tokens $\mathbf{v}$ close to geometric features $\bar{\mathbf{g}}$, it drifts away from the original alignment relative to language tokens $\mathbf{w}$. We posit that this could lead to degradations in pretrained capabilities that jointly use vision $\mathbf{v}$ and language $\mathbf{w}$, as their associations would become out-of-distribution.

Formalization. We formalize the above discussion by letting $\mathbf{v}, \mathbf{w}, \bar{\mathbf{g}} \in \mathbb{R}^{N\times d}$ denote pretrained vision, language, and geometric teacher features with ℓ_2-normalized rows [2]. Holding $\mathbf{w}$ and $\bar{\mathbf{g}}$ fixed, define *alignment* $f(\mathbf{v}, \mathbf{w}) := \sum_{i=1}^{N} \mathbf{v}_i^\top \mathbf{w}_i$ and initial gaps $\delta := \|\mathbf{v} - \bar{\mathbf{g}}\|$ and $\epsilon := \|\mathbf{v} - \mathbf{w}\|$ using the Frobenius norm. Feature distillation minimizes $\mathcal{L}_{\mathsf{f}}(\mathbf{x}) := \ell_{\mathsf{f}}(\mathbf{x}, \bar{\mathbf{g}})$ for a distance metric ℓ_{f}.

Assumption 1. *Motivated by the separate pretraining of the VLM and geometric teacher, we assume* $\delta > a\epsilon$ *for some* $a > 3$.

Theorem 1. *Under Assumption 1, feature distillation has minimizer* $\mathbf{v}_{\mathrm{f}}^* = \bar{\mathbf{g}}$ *and decreases vision–language alignment:*

$$f(\mathbf{v}, \mathbf{w}) - f(\mathbf{v}_{\mathrm{f}}^*, \mathbf{w}) \geq \tfrac{1}{2}\delta(\delta - 2\epsilon) > 0. \tag{2}$$

The proof is in Appendix A.

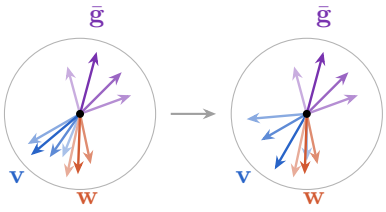
Demonstration. We show this failure mode in a large-scale VLM: LLaVA-Video-7B-Qwen2 [16], by comparing SFT on the VLM3R dataset [1] with SFT combined with feature distillation following 3DRS [2], using the representations from VGGT [48], a large geometry-grounded vision model. For evaluations, we use widely adopted benchmark VSI-Bench [19], categorizing its tasks by the primary type of spatial reasoning required: **(i) geometric-centered** tasks (*Abs. Dist, Room Size, Rel. Dist, Route Plan*), which require aggregating multi-view geometry into an allocentric scene model to reason over inter-object or scene-level relations; **(ii) linguistic grounding-centered** tasks (*Obj. Count, Obj. Size, Appr. Order*), which require precise grounding of text-queried object categories for object localization or commonsense reasoning. We leave an extended discussion in Appendix C.

In this setting, in Table 1 we find that while feature distillation improves geometric tasks, consistent with prior findings [2, 52, 53], it can be *detrimental* to tasks in the second category, which rely on joint use of vision and language such as in per-frame object localization. This implies, like in our toy example, directly grounding features induces a tradeoff in vision-language alignment. The issue was undiagnosed in prior work [2, 52, 53] due to their focus on geometry-centric evaluations, such as 3D object grounding or cross-view relation reasoning, which excluded language-semantic tasks.

To understand this problem at the representation level, in Figure 2(a) we examine the vision-language alignment of the distilled model via cosine similarity scores between vision tokens, and language tokens that refer to a particular object. We observe that, while the vision tokens are grounded to object semantic in SFT, feature distillation indeed loses it, with high scores focused on unrelated but geometrically correlated regions. We further quantify such misalignment by introducing a metric based on language-based object detection precision, comparing the highest-score patches against gold object bounding boxes from ScanNet [58]. Further details of this metric are given in Appendix B.1. Figure 2(b) confirms that the disruption of vision-language alignment happens consistently. Similar representation disruption caused by distillation from off-the-shelf teachers has also been observed in prior work across other domains [29–32]. Our finding here motivates the key question: *How can geometric knowledge be distilled into VLMs without compromising vision-language alignment?*

4 Multi-view relational distillation (MVRD)

We propose multi-view relational distillation (MVRD) to address the challenge raised in Section 3. In MVRD, instead of distilling features themselves, we distill similarities of features across multiple views of a scene. Our hypothesis is that (i) these relations encode geometric correspondences adequate to improve spatial understanding of VLMs, while (ii) allowing the features themselves to remain close to its pretrained vision-language space.



Motivation. The first part of the hypothesis is motivated by studies on human vision [33–36], which suggest that spatial reasoning relies on relational information across a stream of visual inputs. The second part is mathematical, and to illustrate it we recall the two-dimensional example from Section 3.2 with vision, language, and geometric tokens ($\mathbf{v}$, $\mathbf{w}$, $\bar{\mathbf{g}}$), respectively. Instead of matching $\mathbf{v} \approx \bar{\mathbf{g}}$ directly, we consider matching pairwise cosine similarities of $\mathbf{v}$ to that of $\bar{\mathbf{g}}$. Importantly, this leaves vision features *underdetermined*, that is, if $\mathbf{v}^*$ is optimal after distillation, then *every* $t(\mathbf{v}^*)$, where t is any shared rotation and independent scaling of vectors, must be also optimal, as they have the same cosine similarities. This means a learner can choose from these equally geometry-grounded visual representations, favorably one close to its pretrained space, where the alignment between vision $\mathbf{v}$ and language $\mathbf{w}$ is better retained.

Formally, in this motivation, matching pairwise cosine similarities of the row-normalized features in Section 3.2 gives $\mathcal{L}_{\mathrm{r}}(\mathbf{x}) := \ell_{\mathrm{r}}(\mathbf{x}\mathbf{x}^\top, \bar{\mathbf{g}}\bar{\mathbf{g}}^\top)$, where ℓ_{r} is a distance metric on Gram matrices. This loss is invariant to orthogonal transformations (Lemma 1), admitting equally optimal feature orientations. Among these optima, we assume the learner selects one close to its pretrained initialization, motivated by the simplicity bias of gradient-based optimization [59].

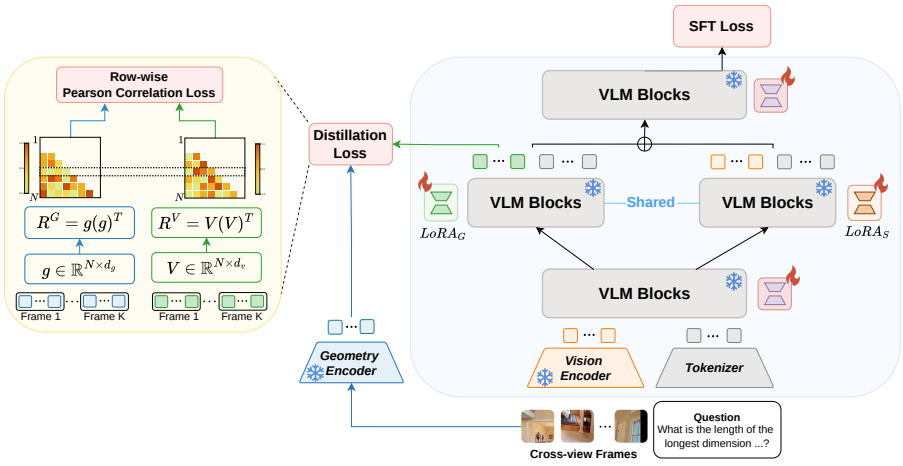


Figure 3: An overview of multi-view relational distillation loss and dual pathway.

Assumption 2. *The distillation procedure selects a relational optimum* $\mathbf{v}_\mathsf{r}^*$ *close to the pretrained initialization* $\mathbf{v}$*, satisfying*

$$\mathcal{L}_\mathsf{r}(\mathbf{v}_\mathsf{r}^*) = 0, \qquad \rho := \|\mathbf{v}_\mathsf{r}^* - \mathbf{v}\| < \epsilon. \tag{3}$$

Recall from Section 3.2 that $\epsilon = \|\mathbf{v} - \mathbf{w}\|$ *denotes the pretrained vision-language gap.*

Theorem 2. *Under Assumptions 1 and 2, relational distillation retains better vision–language alignment than feature distillation:*

$$f(\mathbf{v}_\mathsf{r}^*, \mathbf{w}) > f(\mathbf{v}_\mathsf{f}^*, \mathbf{w}). \tag{4}$$

The proof is in Appendix A.

Method. From these insights, we construct our method as follows. Its core component is *relational representation* $R : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times N}$, defined upon a choice of similarity function $r : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $R(\mathbf{h})[i, j] \coloneqq r(\mathbf{h}[i], \mathbf{h}[j])$. For the vision $\mathbf{v}_{1:T}^l$ and geometric tokens $\bar{\mathbf{g}}_{1:T}$, it defines a multi-view relational representation, with $N = T \times H_p \times W_p$, the latter encoding geometric correspondences, and the goal being its distillation. We therefore define the multi-view relational distillation loss as

$$\mathcal{L}_{\mathsf{RD}} \coloneqq \ell_{\mathsf{RD}}\left(R\left(f(\mathbf{v}_{1:T}^l)\right), R(\mathsf{sg}(\bar{\mathbf{g}}_{1:T}))\right), \tag{5}$$

where $f : \mathbb{R}^d \to \mathbb{R}^{d_g}$ is a position-wise MLP projection, and ℓ_{RD} is a suitable loss function on pairs of $N \times N$ matrices. The VLM and the MLP projector are trained with a composite loss $\mathcal{L}_{\mathsf{SFT}} + \gamma \mathcal{L}_{\mathsf{RD}}$ with hyperparameter $\gamma > 0$. It trains VLM on visual question answering, while making its representations more geometric, in the multi-view relational sense determined by the choice of r.

Choice of similarity. Our motivating example suggests a natural design criterion for the similarity function r. If more features share the same pairwise similarities, there is more freedom to select visual features that remain close to the pretrained space. Cosine similarity is suitable in this regard because it is invariant to feature magnitudes and global orientation of the feature set. Thus, relational distillation only need match relative angles, leaving global orientation and per-feature scales largely unconstrained. By contrast, negative Euclidean distance lacks scale invariance, forcing distillation to also match feature scales. Appendix E.3 verifies that cosine similarity performs better than Euclidean.

Choice of matrix loss. In prior work on relational distillation, mean squared error (MSE) has been used as the loss ℓ_{RD} on relational representation matrices. We empirically find it better to use negative Person correlation computed row-wise and averaged. This is partially since the resulting loss ℓ_{RD} is invariant to the scales of the similarity scores, which improves its robustness across scales, reducing the need to carefully tune the hyperparameters for training stability. In addition to this choice, to adhere to causal attention of VLMs, we apply causal masking before computing the loss, as well as zeroing the diagonal to remove the role of trivial self-similarity. Appendix E.3 verifies that the correlation loss performs better than MSE.

Architectural aspects. To further minimize the drift from the pretrained representation space, we use low-rank adaptation (LoRA) [60]. Yet, as motivated in the beginning of the section, jointly supervising a shared parameter set with $\mathcal{L}_{\mathsf{SFT}}$ and $\mathcal{L}_{\mathsf{RD}}$ may still perturb the language-relevant subspace. We therefore decouple the two objectives through a lightweight parameter separation we refer to as **Dual Pathway** (Figure 3): at several intermediate layers, a *geometric adapter*, a LoRA module supervised by the composite loss, learns from multi-view relational signal, while a parallel *semantic adapter* supervised only by $\mathcal{L}_{\mathsf{SFT}}$ preserves language-aligned behavior. Their outputs are summed after the distillation layer and passed through the remaining VLM blocks. With only modest parameter and latency overhead, this design nearly matches the feature-fusion baseline, as shown in the next section.

5 Experiments

5.1 Main evaluation: Visual-spatial reasoning

Datasets and evaluation. We train on the spatial reasoning dataset released by VLM-3R [1], which contains over 200,000 question–answer pairs grounded in scenes from ScanNet [58], ScanNet++ [61], and ARKitScenes [62]. This dataset has been shown to be effective for learning spatial reasoning [1, 63, 26, 46] while remaining tractable enough to support ablations across diverse settings. We evaluate on VSI-Bench [19], which comprises roughly $5{,}000$ QA pairs spanning eight spatial reasoning tasks, and follow its official evaluation protocol for both Multiple-Choice and Numerical-Answer formats. For both training and evaluation, we sample 32 frames per scene video as multi-view input. To mitigate overfitting when distilling from a geometric teacher tied to fixed frame indices, we apply a frame-sampling augmentation detailed in Appendix D.1.

Baselines. We compare against three groups of models. **(1)** *Proprietary VLMs*, including GPT-4o [38] and Gemini-2.5-Pro [37]. **(2)** *Open-source general-purpose VLMs*, including InternVL3 [15], LLaVA-Video [16], LLaVA-OneVision [64], and Qwen2.5-VL [14]. **(3)** *Spatial Reasoning VLMs*, including VLM-3R [1], VG-LLM [26], 3DRS [2], and LLaVA-Video [16]. For the last category, we only include models fine-tuned on the same training data as ours, which allows us to better isolate methodological contributions from the effect of additional data. Also, all baselines except VG-LLM are based on the same VLM backbone.

Implementation details. Our model is based on LLaVA-Video-7B-Qwen2 [16] (LLaVA-Video-7B hereafter), with VGGT [48] as the frozen geometric teacher. Distillation is performed at 20th layer of VLM backbone (0-indexed; 28 layers total); our dual-pathway variant splits LoRA [60] adapters into geometric and semantic pathways over layers 13–20. We fine-tune the LoRA adapters (rank 128, $\alpha = 256$), the input projection of the vision tokens in VLM, and the distillation MLP while keeping the rest frozen. Training runs on 8 NVIDIA H200 GPUs. Full hyperparameters are in Appendix D.1.

Comparison on VSI-Bench. We evaluate the spatial reasoning capability of our model on VSI-Bench [19] in Table 2. Comparison with feature distillation - 3DRS [2], MVRD improves on 6 of 8 VSI-Bench tasks. The gains are most consistent on the linguistic-grounding task group, where MVRD wins on every task and 3DRS are considerably below the supervised fine-tuned (SFT) LLaVA-Video baseline. This indicates that our relational distillation target reduces vision–language misalignment, suffers from a smaller performance drop on tasks that rely on linguistic object localization, where prior feature distillation degrades the underlying VLM.

At the same time, our relational target injects enough geometric structure to clearly surpass the SFT baseline on three geometric-centered tasks: Abs. Dist., Room Size, and most prominently Rel. Dir., where MVRD exceeds SFT by **+11.1%** and 3DRS* by **+5.5%**. Aggregated over all tasks, MVRD reaches an average accuracy of 59.9%, surpassing the feature-fusion method VG-LLM-8B [26], when trained on the same data.

Moreover, decoupling the geometric and semantic pathways (MVRD + Dual Pathways) further preserves vision–language alignment: accuracy on Appr. Order and Obj. Size closely tracks (and on Obj. Size slightly exceeds) the SFT model. This trades a small drop on geometric-centered tasks for a more balanced overall performance, yielding an average of **60.4%**, within 0.5 points of the state-of-the-art feature-fusion method VLM-3R [1] trained on the same data. As shown in Figure 1, this comes with only modest additional parameters and runtime overhead relative to VLM-3R.

Computation analysis. We examine the inference-time overhead of MVRD + Dual Pathway relative to feature-fusion methods that rely on an auxiliary geometric encoder at test time: VLM-3R [1]

Table 2: **Evaluation on VSI-Bench.** dark green and light green indicate the first and second best open-source results per task column, respectively. Within models' average accuracy, dark purple and light purple are used to denote first and second best performance. **MVRD + Dual Pathway** approaches feature-fusion VLM-3R without using additonal geometric encoder.

Methods	Geo. Enc.	Avg.	Abs. Dist.	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Obj. Count	Obj. Size	Appr. Order
			Geometric					Linguistic grounding		
Baselines										
Human [19]	–	79.2	47.0	45.9	94.7	95.8	95.8	94.3	60.4	100.0
Random Choice	–	–	–	–	25.0	47.9	28.4	–	–	25.2
Proprietary VLMs										
GPT-4o [38]	✗	34.0	5.3	38.2	37.0	41.3	31.5	46.2	43.8	28.5
Gemini-2.5 Pro [37]	✗	51.5	34.9	42.8	61.1	47.8	45.9	43.8	64.3	71.3
Open-source general-purpose VLMs										
Qwen2.5-VL-7B-Instruct [14]	✗	29.3	10.5	29.6	38.4	38.0	29.8	25.2	36.4	26.8
LLaVA-OneVision-7B [64]	✗	32.4	20.2	12.3	42.5	35.2	29.4	47.7	47.4	24.4
LLaVA-Video-7B [16]	✗	35.6	14.0	24.2	43.5	42.4	34.0	48.5	47.8	30.6
LLaVA-OneVision-72B [64]	✗	40.2	23.9	37.5	42.5	39.9	32.5	43.5	57.6	44.6
LLaVA-Video-72B [16]	✗	40.9	22.8	35.3	42.4	36.7	35.0	48.9	57.4	48.6
InternVL3-8B [15]	✗	42.1	39.0	33.6	48.3	36.4	27.3	68.1	48.4	35.4
InternVL3-78B [15]	✗	48.4	53.7	39.5	55.9	39.5	28.9	71.2	44.4	54.5
Spatial Reasoning VLMs										
VG-LLM-8B [26]	✓	58.2	53.8	62.1	63.8	83.0	44.3	71.7	68.8	18.4
VLM-3R [1]	✓	60.9	49.4	67.1	65.4	80.5	45.4	70.2	69.2	40.1
LLaVA-Video-7B (*Fine-tuned*) [16]	✗	57.7	43.6	63.7	64.9	68.9	40.7	70.6	70.8	38.5
3DRS* [2]	✗	57.0	51.3	69.1	65.8	74.5	37.6	70.0	68.0	19.6
MVRD	✗	59.9	50.5	69.3	62.7	80.0	40.7	72.9	68.3	35.1
MVRD + Dual Pathway	✗	60.4	49.6	68.6	64.4	78.2	42.8	71.8	70.1	37.7

Table 3: Ablation on distillation target and dual pathway. *Feat Dist* means feature distillation.

Method	**Distillation Target**	**Trainable Params**	**Dual Path**	**Avg.**	Abs. Dist.	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Obj. Count	Obj. Size	Appr. Order
					Geometric					Linguistic grounding		
3DRS*	feature	Full	–	57.0	51.3	69.1	65.8	74.5	37.6	70.0	68.0	19.6
Feat Dist	feature	LoRA	✗	58.5	48.3	68.4	61.4	75.7	44.9	71.4	67.8	30.1
MVRD	relation	LoRA	✗	59.9	50.5	69.3	62.7	80.0	40.7	72.9	68.3	35.1
Feat Dist + Dual Pathway	feature	LoRA	✓	59.3	49.0	61.7	67.3	76.4	43.8	71.3	69.5	35.8
MVRD + Dual Pathway	relation	LoRA	✓	60.4	49.6	68.6	64.4	78.2	42.8	71.8	70.1	37.7

(CUT3R [49]) and VG-LLM [26] (VGGT [48]). Under matched inputs of 32 views on a single H200 GPU, Figure 1 shows that dual pathway introduces only modest cost compared to integrating a full geometric encoder. Specifically, MVRD + Dual Pathway adds 0.19B parameters - roughly **25%** of CUT3R and **15%** of VGGT - and incurs only **16%** of CUT3R's and **5%** of VGGT's added latency. Combined with the accuracy reported in Table 2, this shows that MVRD + Dual Pathway approaches feature-fusion VLM-3R's performance on VSI-Bench at a fraction of the inference overhead.

5.2 Analysis

Distillation loss. For deeper analysis on improvement, we fix all training settings and compare feature distillation with relation distillation. As shown in Table 3, MVRD consistently improves the average score over the corresponding feature distillation baseline, both with and without dual pathway. It suggests that relation distillation provides a stronger geometric learning signal while better preserving vision-language alignment. We also find that LoRA-based feature distillation outperforms the fully fine-tuned 3DRS* [2], indicating that LoRA helps mitigate catastrophic forgetting.

Table 4: Evaluation on VSI-Bench and additional general VQA benchmarks.

Method	Distillation Target	Trainable Params	Dual Path	Avg.	VSI-Bench	BLINK*	MuirBench*
Feat Dist	feature	LoRA	✗	45.5	58.5	36.8	41.1
MVRD	relation	LoRA	✗	47.3	59.9	40.3	41.6
Feat Dist + Dual Pathway	feature	LoRA	✓	47.6	59.3	39.8	43.7
MVRD + Dual Pathway	relation	LoRA	✓	48.3	60.4	40.9	43.6

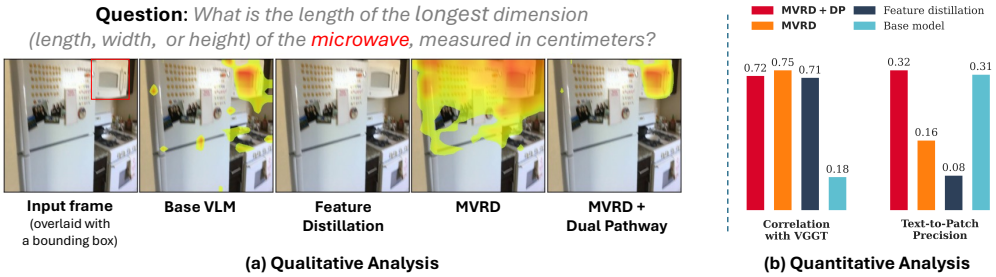


Figure 4: **Vision-language alignment across methods. (a)** At distillation layer, MVRD and MVRD + Dual Pathway align *microwave* with its visual patches better than feature distillation. **(b)** Both improve vision-language alignment while matching feature distillation in geometric relation modeling.

Evaluation on general VQA benchmarks. To evaluate transfer of MVRD to general visual question answering (VQA), we extend the controlled comparison (Table 3) to the multi-image subsets of BLINK [65] and MuirBench [66] (denoted BLINK* and MuirBench*). These benchmarks assess understanding across multiple images through tasks such as art-style matching, forensic detection, allowing us to examine transfer beyond multi-view scene geometry. As shown in Table 4, which also includes VSI-Bench, MVRD improves all three scores over feature distillation without Dual Pathway, raising the average from 45.5 to 47.3. Dual Pathway further improves all three scores for both distillation targets. With Dual Pathway, MVRD outperforms feature distillation on VSI-Bench and BLINK*, while scoring 0.1 points lower on MuirBench*, yielding a higher overall average. These results demonstrate benefits of both relational distillation and Dual Pathway beyond VSI-Bench.

Analysis on visual representation. To better understand the representation space induced by distillation, we provide a qualitative example in Figure 4(a); additional multi-frame examples are deferred to Appendix E.2. At the distillation layer, feature distillation visibly degrades text–vision alignment: object text tokens assign low cosine similarity to their corresponding visual patches. In contrast, MVRD yields high similarity to both the corresponding patch and its neighbors. We attribute this to relation learning from the geometric teacher: since local patches encode local 3D regions and the teacher's multi-view relations reflect physical-world proximity (Appendix B.4, Figure 5), the VLM's patch features inherit this proximity, spreading text–patch similarity to neighboring regions. MVRD + Dual Pathway resolves this by decoupling multi-view relation learning from the pre-trained vision–language alignment, yielding precise text–patch correspondence.

We further quantify this effect along two complementary axes, both computed at the distillation layer on a subset of ScanNet validation scenes. *(i)* geometric relation is measured by $-\mathcal{L}_{\text{RD}}$, i.e. the row-wise Pearson correlation between VLM's relation representation $R(\mathbf{v}^{l}_{1:T})$ and VGGT's relation representation $R(\bar{\mathbf{g}})$; *(ii)* Vision–language alignment is measured by text-grounded object detection precision. As shown in Figure 4(b), both distillation approaches capture geometric relations, but MVRD better preserves vision-language alignment than feature distillation, achieving 0.16 vs. 0.08 detection precision. With efficient parameter separation, MVRD + Dual Pathway achieves both properties, explaining its leading average accuracy among distillation settings in Table 2.

Distillation and dual pathway layers. We ablate the choice of distillation layer in Table 5. The results show that distillation from early or middle layers underperforms the SFT-only baseline, suggesting that the VLM requires sufficient computational depth to capture geometric relations. In contrast, later-layer distillation improves performance, with the 20th layer achieving the best overall score. Also, Table 6 shows the impact of parameter separation layer in our dual pathway design: applying separation to layers 13–20 yields the best result among the evaluated choices.

Table 5: Distillation Layer Ablation.

Distill. Layer	L7	L14	L20	Last	SFT-only
Overall Score	53.9	57.3	59.9	59.8	57.7

Table 6: Separation Layer Selection Ablation.

Layers for Sep.	0–20	7–20	10–20	13–20	16–20
Overall Score	60.2	59.8	60.0	60.4	59.5

Table 7: Performance comparison across different VLM backbones.

Training Method	LLaVA-Video-7B	InternVL3-8B	Qwen2.5-VL-7B
Base (no fine-tuning)	35.6	42.1	29.3
SFT	57.7	50.3	56.1
Feature Distillation	58.5	57.0	55.5
MVRD	59.9	58.7	56.1
MVRD + Dual Pathway	60.4	59.9	59.8

Table 8: Results on 3D scene understanding

Method	3D Encoder	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		SQA3D
		Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	C@0.5	B-4@0.5	C	EM	EM
Geometry-based Methods										
Grounded 3D-LLM [67]	✓	47.9	44.1	45.2	40.6	70.6	35.5	72.7	-	-
PQ3D [68]	✓	57.0	51.2	-	50.1	80.3	36.0	-	-	47.1
ChatScene [69]	✓	55.5	50.2	57.1	52.4	77.1	36.3	87.7	21.6	54.6
Inst3D-LLM [70]	✓	57.8	51.6	58.3	53.5	79.7	38.3	88.6	24.6	-
3D-LLaVA [71]	✓	51.2	40.6	-	-	78.8	36.9	92.6	-	54.5
View-based Methods										
LLaVA-3D [72]	✗	54.1	42.4	-	-	79.2	41.1	91.7	27.0	55.6
Video-3D LLM [73]	✗	58.1	51.7	58.0	52.7	83.8	41.3	102.1	30.1	58.6
Ross3D [74]	✗	61.3	54.7	59.4	54.2	85.6	47.5	107.3	31	60.9
3DRS [2]	✗	62.9	56.1	60.4	54.9	86.1	41.6	104.8	30.3	60.6
MVRD	✗	63.3	56.4	61.1	55.6	86.7	42.2	104.9	30.4	60.8

Robustness on VLM backbone. We additionally evaluate MVRD on two representative VLM backbones, InternVL3-8B [15] and Qwen2.5-VL-7B [14], using the same training setup as in Section 5.1 except for the backbone. Under the controlled setting, as shown in Table 7, MVRD consistently outperforms feature distillation with further gains from Dual Pathway, consistent with the previous observation. It shows that the proposed relational distillation is robustly beneficial across diverse VLM backbones. More detailed results are presented in Table 13.

3D scene understanding. To test whether MVRD generalizes beyond geometry-centric spatial QA, we train a separate model under the 3DRS scene-understanding protocol [2]. This setting uses five standard 3D scene-understanding datasets: ScanRefer [75], Multi3DRefer [76], Scan2Cap [77], ScanQA [78], and SQA3D [79], covering object grounding, dense captioning, and 3D question answering. Details are provided in Appendix D.2. Table 8 shows that MVRD transfers well to 3D scene understanding tasks. It achieves the strongest results on object grounding, outperforming prior 3D scene understanding generalists on ScanRefer and Multi3DRefer, and improves over the feature-distillation baseline 3DRS [2] across all metrics. These gains are consistent with our analysis in Section 5.2: MVRD injects multi-view geometric structure while better preserving text–visual alignment, which is especially important for grounding.

6 Conclusion

We addressed the problem of teaching vision-language models geometry without inflating inference cost or eroding pretrained vision-language alignment. We first identified a failure mode of feature distillation from geometry-grounded teachers: matching multi-view features pulls student representations away from the pretrained vision-language space, degrading linguistic grounding such as object-order tracking. We proposed **multi-view relational distillation (MVRD)**, which distills patch-wise cosine similarities across views, which can satisfy the geometric objective while staying near its pretrained representation. Across representative VLMs, MVRD improves VSI-Bench accuracy over strong baselines, narrows the gap to a state-of-the-art feature fusion baseline, and transfers to 3D object grounding, dense captioning, and question answering.

Limitations. Our study is conducted under moderate-scale training, and the method still induces some trade-offs. Future work could address this with large-scale training paired with efficient architectures that can encode both semantics and geometry.

Acknowledgments This work was in part supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) (RS-2022-II220926, RS-2026-25617720, and RS-2019-II190075) funded by the Korean government (MSIT). We also thank Phillip Y. Lee and Duy M. H. Nguyen for their thoughtful discussions.

References

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

[2] 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.

[3] Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1967.

[4] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982. ISBN 0716715678.

[5] Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

[6] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[7] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

[8] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978.

[9] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[11] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[12] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[13] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6585–6597, 2025.

[14] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[15] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[16] 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*, 2025. ISSN 2835-8856.

[17] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024.

[18] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Stibench: Are mllms ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5622–5632, 2025.

[19] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[20] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8600–8612, 2025.

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

[22] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2025.

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

[24] Akarsh Kumar, Jeff Clune, Joel Lehman, and Kenneth O Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*, 2025.

[25] Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G2vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. *arXiv preprint arXiv:2511.21688*, 2025.

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

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

[28] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[29] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *Advances in Neural Information Processing Systems*, 38:122647–122676, 2026.

[30] Aritra Bhowmik, Denis Korzhenkov, Cees G Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. In *International Conference on Learning Representations*, volume 2026, pages 122379–122397, 2026.

[31] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

[32] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23922–23932. IEEE, 2024.

[33] Nuo Li and James J DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *science*, 321(5895):1502–1507, 2008.

[34] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

[35] James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

[36] Luca Tommasi and Bruno Laeng. Psychology of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(6):565–580, 2012.

[37] 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.

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

[39] Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Stibench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025.

[40] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. *arXiv preprint arXiv:2508.02095*, 2025.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[42] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

[43] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

[44] Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

[45] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2026.

[46] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

[47] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025.

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

[49] 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)*, pages 10510–10522. IEEE, 2025.

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

[51] Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. *arXiv preprint arXiv:2511.21688*, 2025.

[52] Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2026.

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

[54] Peter Hugoe Matthews. *A short history of structural linguistics*. Cambridge University Press, 2001.

[55] Tom Leinster. *Basic category theory*. Cambridge University Press, 2014.

[56] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[57] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[58] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2432–2443. IEEE, 2017.

[59] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[60] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[61] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. *arXiv preprint arXiv:2308.11417*, 2023.

[62] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2022.

[63] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2026.

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

[65] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[66] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pages 62624–62650, 2025.

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

[68] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. *arXiv preprint arXiv:2405.11442*, 2024.

[69] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2024.

[70] Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. *arXiv preprint arXiv:2503.00513*, 2025.

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

[72] 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*, 2025.

[73] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. *arXiv preprint arXiv:2412.00493*, 2025.

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

[75] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *arXiv preprint arXiv:1912.08830*, 2020.

[76] Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. *arXiv preprint arXiv:2309.05251*, 2023.

[77] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. *arXiv preprint arXiv:2012.02206*, 2020.

[78] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. *arXiv preprint arXiv:2112.10482*, 2022.

[79] 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*, 2023.

[80] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2023.

A Proofs

Proof of Theorem 1. Since ℓ_{f} is a distance metric, its minimum is zero and is attained if and only if the student feature equals the teacher feature, that is $\mathbf{v}_{\mathsf{f}}^* = \bar{\mathbf{g}}$, which proves the first part. For the second part, the Frobenius norm property $\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 - 2f(\mathbf{x},\mathbf{y}) + \|\mathbf{y}\|^2$ gives

$$\|\bar{\mathbf{g}}-\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2 = \|\bar{\mathbf{g}}\|^2 - \|\mathbf{v}\|^2 + 2\big(f(\mathbf{v},\mathbf{w}) - f(\bar{\mathbf{g}},\mathbf{w})\big). \tag{6}$$

Since each row of $\bar{\mathbf{g}}$ and $\mathbf{v}$ is ℓ_2-normalized, $\|\bar{\mathbf{g}}\| = \|\mathbf{v}\| = \sqrt{N}$, so

$$f(\mathbf{v},\mathbf{w}) - f(\bar{\mathbf{g}},\mathbf{w}) = \tfrac{1}{2}\big(\|\bar{\mathbf{g}}-\mathbf{w}\|^2 - \epsilon^2\big). \tag{7}$$

By the triangle inequality, $\|\bar{\mathbf{g}}-\mathbf{w}\| \geq \delta - \epsilon$, hence

$$f(\mathbf{v},\mathbf{w}) - f(\bar{\mathbf{g}},\mathbf{w}) \geq \tfrac{1}{2}\big((\delta-\epsilon)^2 - \epsilon^2\big) = \tfrac{1}{2}\delta(\delta-2\epsilon) > 0, \tag{8}$$

using $\delta > 3\epsilon$ and $\mathbf{v}_{\mathsf{f}}^* = \bar{\mathbf{g}}$. □

Lemma 1 (Orthogonal invariance of relational distillation)**.** *For any $\mathbf{x} \in \mathbb{R}^{N\times d}$ and any orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d\times d}$,*

$$\mathcal{L}_{\mathsf{r}}(\mathbf{x}) = \mathcal{L}_{\mathsf{r}}(\mathbf{x}\mathbf{Q}). \tag{9}$$

Proof. Orthogonality gives

$$(\mathbf{x}\mathbf{Q})(\mathbf{x}\mathbf{Q})^\top = \mathbf{x}\mathbf{Q}\mathbf{Q}^\top\mathbf{x}^\top = \mathbf{x}\mathbf{x}^\top. \tag{10}$$

Since $\mathcal{L}_{\mathsf{r}}$ depends on $\mathbf{x}$ only through its Gram matrix, its value is unchanged. □

For any relational minimizer $\mathbf{v}_{\mathsf{r}}$, Lemma 1 gives $\mathcal{L}_{\mathsf{r}}(\mathbf{v}_{\mathsf{r}}\mathbf{Q}) = \mathcal{L}_{\mathsf{r}}(\mathbf{v}_{\mathsf{r}}) = 0$ for every orthogonal $\mathbf{Q}$. These transformations also preserve row norms, yielding equally optimal feasible representations with different global orientations. Assumption 2 further requires the selected optimum to lie close to $\mathbf{v}$. The alignment comparison below uses this additional assumption.

Proof of Theorem 2. By Assumption 2, the distillation procedure selects an optimum $\mathbf{v}_{\mathsf{r}}^*$ satisfying $\rho = \|\mathbf{v}_{\mathsf{r}}^* - \mathbf{v}\| < \epsilon$. Since the feature rows are ℓ_2-normalized, $\|\mathbf{v}_{\mathsf{r}}^*\| = \|\mathbf{v}\| = \sqrt{N}$. The Frobenius norm identity therefore gives

$$f(\mathbf{v},\mathbf{w}) - f(\mathbf{v}_{\mathsf{r}}^*,\mathbf{w}) = \tfrac{1}{2}\big(\|\mathbf{v}_{\mathsf{r}}^* - \mathbf{w}\|^2 - \epsilon^2\big). \tag{11}$$

By the triangle inequality, $\|\mathbf{v}_{\mathsf{r}}^* - \mathbf{w}\| \leq \rho + \epsilon$, so

$$f(\mathbf{v},\mathbf{w}) - f(\mathbf{v}_{\mathsf{r}}^*,\mathbf{w}) \leq \tfrac{1}{2}\rho(\rho+2\epsilon). \tag{12}$$

Combining this with Theorem 1 yields

$$f(\mathbf{v}_{\mathsf{r}}^*,\mathbf{w}) - f(\mathbf{v}_{\mathsf{f}}^*,\mathbf{w}) \geq \tfrac{1}{2}\delta(\delta-2\epsilon) - \tfrac{1}{2}\rho(\rho+2\epsilon) \tag{13}$$

$$= \tfrac{1}{2}(\delta+\rho)(\delta-\rho-2\epsilon). \tag{14}$$

Under Assumption 1, $\delta > 3\epsilon$. Together with $\rho < \epsilon$, this implies $\delta - \rho - 2\epsilon > \epsilon - \rho > 0$, proving the claim. □

B Implementation Details

B.1 Text-based Object Detection Precision

We use text-based object detection precision as a quantitative probe of vision–language alignment in the VLM's internal visual representation. We select visual patches with high cosine similarity to the object text token and measure the fraction that belong to the ground-truth patch set. The metric is reported in Figure 2(b) and Figure 4(b) to assess the vision-language alignment of the distilled and base models.

Question selection. We restrict the probe to object nouns from the *Object Size* task of VSI-Bench [19], as each question refers to a single object class under a predefined format, making object noun extraction straightforward. We further keep only those questions whose underlying scene video originates from ScanNet [58], so that ground-truth 3D bounding boxes, camera poses, and depth maps are all available. To investigate the performance drop caused by distillation, we select questions on which SFT outperforms the distillation methods. This yields 42 questions from 42 scenes.

Forward pass and feature extraction. For each question, we uniformly sample 32 frames from the scene video and feed them, together with the question text, into the VLM. At layer ℓ we extract: (i) the visual-token hidden states $\{H_i^{(\ell)} \in \mathbb{R}^{h_v \times w_v \times d_v}\}_{i=1}^{32}$, and (ii) the hidden states of the subword tokens that spell out the object's name in the prompt, $\{e_j \in \mathbb{R}^{d_v}\}_{j=1}^{T}$. We extract the object name from the fixed template used for VSI-Bench Object Size questions (e.g., the `<object>` slot in *"What is the length of the longest dimension ... of the* `<object>`*... ?"*), and the corresponding subword tokens are located by string matching against the tokenized prompt. We average-pool these object-name token features into a single query vector $\bar{e} = \frac{1}{T}\sum_{j=1}^{T} e_j$, which we refer to as the *noun token*.

Predicted patch set. We compute the cosine similarity $s_{i,u,v} = \cos(\bar{e}, H_i^{(\ell)}[u,v])$ between the noun token and every visual patch across all $32 \times h_v \times w_v$ positions in the scene, and let $s_{\max} = \max_{i,u,v} s_{i,u,v}$. Assuming that the queried object appears in the sampled frames, we define the predicted patch set adaptively as

$$\mathcal{P}_{\text{pred}} \;=\; \big\{(i,u,v) \;:\; s_{i,u,v} \;\geq\; 0.9\, s_{\max}\big\}.$$

This relative threshold avoids the dependence on the scale of cosine similarity scores across different examples and model layers.

Ground-truth patch set. For each frame i, we identify ground-truth object patches by comparing the 3D coordinates extracted for each visual patch against the queried object's ground-truth 3D bounding box. We first unproject each valid depth pixel into world coordinates using the depth map, camera intrinsics K_i, and camera pose/extrinsics provided by ScanNet (Appendix B.2). Pixels are then grouped according to the image patch grid. For each patch, we compute the fraction of its valid unprojected 3D points that fall inside the queried object's 3D bounding box, using coordinate-wise bounds checking against the box's minimum and maximum corners. A patch is labeled as a ground-truth patch if this fraction is at least 50%. The ground-truth patch set $\mathcal{P}_{\text{gt}}$ contains all ground-truth patches across frames.

Precision. For each question we compute

$$\text{Prec} \;=\; \frac{|\mathcal{P}_{\text{pred}} \cap \mathcal{P}_{\text{gt}}|}{|\mathcal{P}_{\text{pred}}|},$$

and report the mean precision across all selected questions in Figure 2, Figure 4.

B.2 3D Back-projection

We use standard pinhole 3D back-projection on ScanNet using the camera parameters and depth maps released with the dataset. For a pixel (u,v) with depth $d_{u,v}$ in a frame with intrinsics $K \in \mathbb{R}^{3\times3}$ and world-to-camera extrinsics $[R \mid t] \in \mathbb{R}^{3\times4}$, the 3D point in world coordinates is

$$\mathbf{p}_{\text{cam}} = d_{u,v} \cdot K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \qquad \mathbf{p}_{\text{world}} = R^{\top}\big(\mathbf{p}_{\text{cam}} - t\big).$$

Pixels with invalid (zero or missing) depth are discarded.

B.3 Pearson Correlation

We construct our relational distillation objective in Eq (5) using negative row-wise Pearson correlation between two relational representation matrices. Given two real-valued vectors $a, b \in \mathbb{R}^n$, the Pearson

correlation coefficient is defined as:

$$\rho(a,b) \;=\; \frac{\sum_{k=1}^{n}(a_k-\bar{a})(b_k-\bar{b})}{\sqrt{\sum_{k=1}^{n}(a_k-\bar{a})^2}\,\sqrt{\sum_{k=1}^{n}(b_k-\bar{b})^2}},$$

where $\bar{a}$ and $\bar{b}$ are the sample means. Pearson correlation is invariant to positive affine rescaling ($\rho(\alpha a+\beta,\, b) = \rho(a,b)$ for $\alpha > 0$), which makes it suitable for comparing two quantities measured on different scales, such as cosine distance and 3D Euclidean distance.

For two square matrices $A, B \in \mathbb{R}^{N\times N}$, we define the *row-wise* Pearson correlation as the average of per-row Pearson correlations,

$$\rho_{\text{row}}(A,B) \;=\; \frac{1}{N}\sum_{i=1}^{N}\rho(A_{i,:},\, B_{i,:}).$$

B.4 Correlation between VGGT Multi-view Relations and 3D Distances

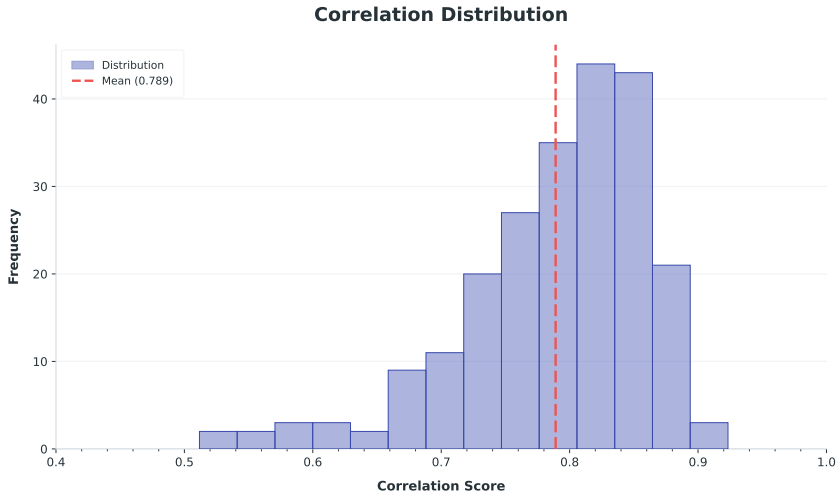


Figure 5: Row-wise Pearson correlation between VGGT cosine distances and 3D Euclidean distances on the ScanNet validation set.

We empirically verify that the multi-view similarity matrix R^G extracted from VGGT [48] captures meaningful 3D structure by comparing it to pairwise 3D Euclidean distances between the same patches. For each scene video in the ScanNet validation set we proceed as follows.

1. **Frame sampling.** Uniformly sample 32 frames from the scene video and resize each to 378×378.
2. **VGGT features.** Pass all 32 frames jointly through VGGT and extract patch-level feature maps of shape $32 \times 27 \times 27 \times d_g$ (patch size 14).
3. **Multi-view similarity.** Flatten the feature maps to $H^G \in \mathbb{R}^{N\times d_g}$ with $N = 32 \cdot 27 \cdot 27 = 23{,}328$, ℓ_2-normalize each row to obtain $\tilde{H}^G$, and compute the Gram matrix $R^G = \tilde{H}^G(\tilde{H}^G)^\top \in \mathbb{R}^{N\times N}$.
4. **3D back-projection.** For each frame, back-project all pixels with valid depth into 3D world coordinates following Appendix B.2, using the depth and camera parameters distributed with ScanNet.
5. **Patch-level 3D centroids.** For each 14×14 patch, average the 3D coordinates of its constituent valid-depth pixels to obtain a single 3D centroid per patch. We retain only patches that contain at least one valid-depth pixel; let $\mathcal{V} \subseteq \{1,\ldots,N\}$ denote the index set of retained patches.

6. **Pairwise 3D distances.** Stack the centroids of the retained patches into $P \in \mathbb{R}^{|\mathcal{V}|\times 3}$ and compute the pairwise Euclidean distance matrix $D \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ with $D_{ij} = \|P_i - P_j\|_2$.
7. **Correlation.** Restrict R^G to the same row/column index set $\mathcal{V}$, treat $1 - R^G_{\mathcal{V},\mathcal{V}}$ as a learned distance, and compute the row-wise Pearson correlation $\rho_{\text{row}}\big(1 - S^G_{\mathcal{V},\mathcal{V}},\, D\big)$ as defined in Appendix B.3.

The mean row-wise Pearson correlation across ScanNet validation scenes is $\mathbf{0.79}$, with the detailed correlation distribution in Figure 5. This result supports using VGGT's multi-view relations as a proxy for 3D proximity and as a distillation target when ground-truth 3D annotations are unavailable.

C Categorization of Spatial Reasoning Tasks

To better diagnose the trade-off observed in our analysis (Tables 1 to 3), we partition the eight VSI-Bench tasks into two categories according to the dominant type of spatial reasoning each one demands:

1. **Geometry-centered tasks** (*Abs. Dist., Room Size, Rel. Dist., Rel. Dir., Route Plan*): tasks that require aggregating observations across frames into a coherent allocentric model of a scene and performing transformations between viewpoints to estimate the geometric relation in the scene.
2. **Linguistic-grounding-centered tasks** (*Obj. Count, Obj. Size, Appr. Order*): tasks whose answers are primarily determined by reliably grounding object categories in input frames, with the scene's geometric structure playing a secondary role.

Tasks in geometry-centered group.

- *Absolute Distance.* The model must combine views to infer the spatial arrangement of the two queried objects, then estimate their distance using commonsense knowledge of physical scale.
- *Room Size.* Aggregate all input views into a global estimate, while there is no requirement for object grounding inside the scene.
- *Relative Distance / Direction.* The model must combine cross-view observations and conduct viewpoint transformation to construct the partial scene that contains all mentioned objects. While linguistic cues specify which objects to focus on, these tasks require extensive viewpoint transformation to compare the distances/directions of different object regions to the reference object.
- *Route Plan.* The model must track changes in position and viewing direction while reasoning about free space and obstacles.

Tasks in the linguistic-grounding-centered group.

- *Object Count.* Requires detecting every instance of a named category and de-duplicating across frames based on visual cues rather than egocentric-allocentric viewpoint transformation.
- *Object Size.* Requires precise object grounding and common-sense priors about object sizes, including the target object and other objects in the scene as references, while the geometric reasoning plays a secondary role.
- *Appearance Order.* Focus on temporal grounding - identify when each object first appears and order the objects by those times - without any need for an allocentric scene model.

The two groups emphasize different capabilities: integrating scene geometry across views, and grounding object names in individual frames while tracking them over time.

D Training Setups

D.1 Spatial Reasoning Experimental Details

Frame-sampling augmentation. Using the same sampled frames throughout training can encourage the student to memorize relations between fixed frame positions rather than learning the underlying scene geometry, especially when the number of scene videos is limited (about 2000 videos in VLM-3R dataset [1]). We mitigate this with a simple augmentation in the frame-sampling scheme. Given a scene video of N frames, we partition the frames into 32 contiguous chunks of equal length $S = \lfloor N/32 \rfloor$. At every training step, we draw an offset r uniformly from the first chunk, $r \sim \mathrm{Uniform}\{0, 1, \ldots, S-1\}$, and use the 32 input frames at indices

$$\big\{\, r,\ r+S,\ r+2S,\ \ldots,\ r+31S \,\big\}.$$

This places exactly one frame in each chunk and preserves the uniform stride along the temporal axis, while varying the visual information across training iterations. We found that a uniform interval between consecutive frames is important, as large temporal gaps introduce abrupt changes and reduce overlapping visual cues for scene understanding. The augmentation is applied only during training; evaluation follows the deterministic uniform sampling.

Hyperparameters. Table 9 summarizes the training configuration used for our spatial-reasoning experiments. Following VLM-3R [1], we use a learning-rate schedule configured for five epochs and stop training after one epoch. Training took 17 hours with 8 NVIDIA H200 GPUs.

Table 9: Training hyperparameters for the spatial-reasoning experiments on the VLM-3R [1] dataset.

Hyperparameter	Value
Distillation	
Distillation loss weight γ	0.5
LoRA adapter	
LoRA enabled	True
Rank r	128
Alpha α	256
Optimization	
Scheduled training epochs	5 (stopped after epoch 1)
Num GPUS	8
Per-device batch size	1
Gradient accumulation steps	16
Peak learning rate	2×10^{-5}
Weight decay	0.0
Warmup ratio	0.03
Learning-rate scheduler	cosine
System	
Numerical precision	BF16 (TF32 enabled)
Maximum sequence length	32,768 tokens
Gradient checkpointing	False

D.2 Scene Understanding Experimental Details

For the scene-understanding results in Table 8, we train a separate model using the 3DRS protocol on a separate mixture of 3D scene-understanding datasets, rather than on the spatial-reasoning training data used for VSI-Bench. This experiment evaluates the applicability of MVRD to object grounding, dense captioning, and 3D question answering.

Datasets. Following 3DRS [2], we train on the union of five standard 3D scene-understanding datasets, covering object grounding, dense captioning, and question answering. ScanRefer [75] evaluates single-object grounding from natural-language descriptions. Multi3DRefer [76] extends

Table 10: Datasets, evaluation splits, and metrics used for the 3D scene-understanding experiments.

Dataset	Task type	Eval split	Metrics
ScanRefer [75]	Single-object grounding	Val.	Acc@0.25, Acc@0.5
Multi3DRefer [76]	Multi-object grounding	Val.	F1@0.25, F1@0.5
Scan2Cap [77]	Dense object captioning	Val.	C@0.5, B-4@0.5
ScanQA [78]	3D question answering	Val.	CIDEr, EM
SQA3D [79]	Situated 3D question answering	Test	EM

Table 11: Training configuration for the 3D scene-understanding experiments.

Hyperparameter	Value
Base model	LLaVA-Video-7B-Qwen2 [16]
Training protocol	3DRS / Video-3D-LLM protocol [2, 73]
Training datasets	ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, SQA3D
Trainable setting	Full fine-tuning
Training epochs	1
Optimizer	Adam
Effective batch size	16
Language-model learning rate	1×10^{-5}
Vision-encoder learning rate	2×10^{-6}
Warmup ratio	0.03
Object proposals	Mask3D [80]
Distillation layer	Last VLM layer
Hardware	8 NVIDIA H200 GPUs

this setting to descriptions that may refer to multiple objects. Scan2Cap [77] evaluates dense captioning for an object specified by its 3D location within a scene. ScanQA [78] evaluates general 3D question answering, while SQA3D [79] evaluates situated question answering, where the model must reason from an embodied viewpoint in a 3D scene.

Evaluation protocol. We follow the evaluation protocol used by prior 3D scene-understanding generalists [73, 2, 74]. We evaluate on the official test split of SQA3D and on the validation splits of ScanRefer, Multi3DRefer, Scan2Cap, and ScanQA. For ScanRefer, we report grounding accuracy at IoU thresholds of 0.25 and 0.5, denoted Acc@0.25 and Acc@0.5. For Multi3DRefer, we report F1 scores at the same IoU thresholds. For Scan2Cap, we report CIDEr and BLEU-4 under the IoU 0.5 criterion, denoted C@0.5 and B-4@0.5. For ScanQA, we report CIDEr and Exact Match (EM). For SQA3D, we report EM as the metric.

Baselines. We compare with two families of 3D scene-understanding generalists. The first family uses dedicated 3D encoders to process point clouds or object-level 3D representations, including Chat-Scene [69], Grounded 3D-LLM [67], PQ3D [68], Inst3D-LMM [70], and 3D-LLaVA [71]. The second family builds on video VLMs augmented with 3D positional information or 3D-aware supervision, including LLaVA-3D [72], Video-3D-LLM [73], Ross3D [74], and 3DRS [2]. When a baseline does not report a metric for a particular benchmark, we leave the corresponding entry blank in the main comparison table.

Object proposals and prediction format. For tasks requiring object-level grounding or captioning, we follow 3DRS [2] and use object proposals extracted by Mask3D [80]. Object-level visual features are obtained by aggregating the visual tokens associated with each 3D proposal. For grounding tasks, predictions are scored using the similarity between text features and proposal-level visual features, following the evaluation setting of prior work.

Training details. We use LLaVA-Video-7B-Qwen2 [16] as the base VLM and follow the full fine-tuning recipe of 3DRS [2] and Video-3D-LLM [73]. The model is trained for one epoch with Adam optimizer, an effective batch size of 16, and a warmup ratio of 0.03. We use a peak learning rate of 1×10^{-5} for the language model and 2×10^{-6} for the vision encoder. For MVRD, multi-view relational distillation is applied at the last layer of the VLM. The model is trained on 8 NVIDIA H200 GPUs. The detailed configuration is shown in Table 11.

E Additional Results

E.1 3DRS Baseline Performance on VSI-Bench

Table 12: Comparison of 3DRS [2] and the SFT baselines [26, 1] on VSI-Bench

Methods	Avg.	*Abs. Dist.*	*Room Size*	*Rel. Dist.*	*Rel. Dir.*	*Route Plan*	*Obj. Count*	*Obj. Size*	*Appr. Order*
		Geometric					**Linguistic grounding**		
Training on VG LLM [26] S1 data									
SFT [14]	**49.8**	**36.0**	**59.7**	**45.8**	38.9	**30.4**	68.6	**57.9**	**60.2**
SFT + feature distillation (3DRS) [2]	45.9	34.8	56.6	40.9	**43.2**	**30.4**	**68.7**	53.6	39.2
Training on VLM-3R [1] data									
SFT [16]	**57.7**	43.6	63.7	64.9	68.9	**40.7**	**70.6**	**70.8**	**38.5**
SFT + feature distillation (3DRS*) [2]	57.0	**51.3**	**69.1**	**65.8**	**74.5**	37.6	70.0	68.0	19.6

We report our reproduced results (Table 1 in the main text) and the original results of 3DRS [2] on VSI-Bench in Table 12. Compared to the SFT baseline using the same training dataset, 3DRS underperforms on most tasks and falls considerably short overall. The original and reproduced results both show lower overall accuracy than their corresponding SFT, confirming the detrimental effect of feature distillation.

E.2 More Qualitative Results

We supplement Figure 4 of the main text with additional qualitative comparisons (Figure 6) of text-to-patch similarity patterns across all four model variants: the base LLaVA-Video, feature distillation, our relational distillation (MVRD), and MVRD + Dual Pathway under the same training setting using LoRA. For each example, we visualize the cosine similarity between the noun token of the queried object and the visual patches across all 32 input frames.

Visualization detail. For visualization, the similarity scores are min–max normalized separately for each scene and each model variant as

$$\tilde{s}_{m,n} = \frac{s_{m,n} - \min_n s_{m,n}}{\max_n s_{m,n} - \min_n s_{m,n}},$$

where $s_{m,n}$ denotes the cosine similarity for model m and patch n within a given scene. This normalization maps the scores of each model to $[0, 1]$, allowing the color scale to highlight the relative spatial distribution of high-similarity patches within each model. Since the normalization is performed independently for each model and scene, the maps should be interpreted as comparisons of localization patterns rather than absolute similarity magnitudes across models. To improve visibility, we mask out the lowest 65% of patch scores for each scene–model pair and visualize only the regions with relatively high text-to-patch similarity.

Observations. Compared with feature distillation, MVRD better preserves the concentration of high-similarity patches around the queried objects, which suggests that MVRD causes less disruption of the pre-trained feature space. Furthermore, MVRD + Dual Pathway also retains the localized similarity patterns of the base model.

E.3 More Quantitative Results

Per-task VSI-Bench results across VLM backbones. Table 13 reports the full per-task breakdown of Table 7 in the main text for InternVL3-8B and Qwen2.5-VL-7B. We compare the Base VLM, supervised fine-tuning (SFT), feature distillation, MVRD, and MVRD + Dual Pathway. Across both backbones, MVRD consistently outperforms *feature distillation* on most of the tasks from both categories: *geometric-centered* and *linguistic grounding-centered* QA, mirroring the trend observed for LLaVA-Video-7B.

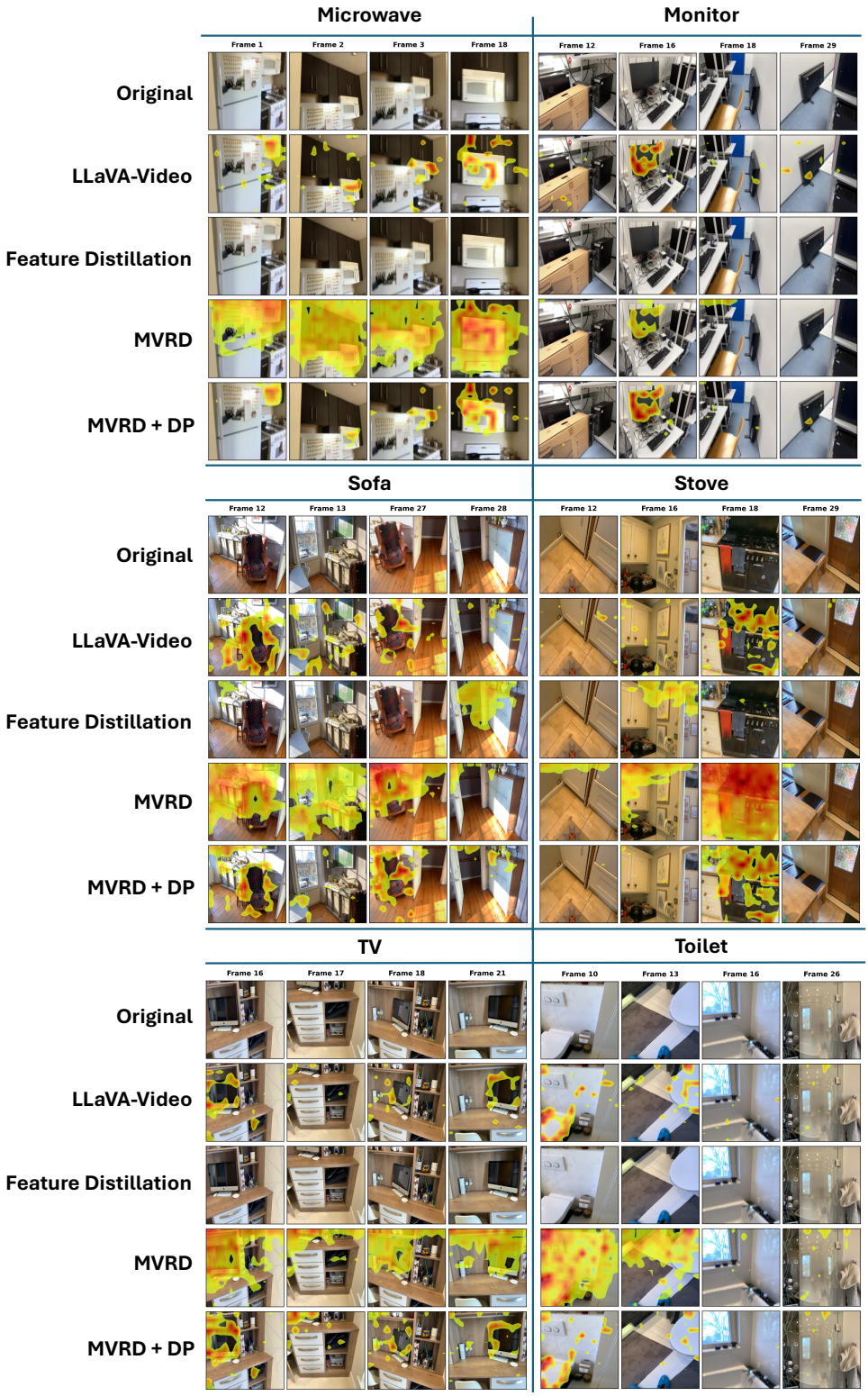


Figure 6: Additional text-to-patch similarity maps. MVRD and MVRD + Dual Pathway retain localized similarity patterns around the queried objects, whereas feature distillation diffuses or misaligns high-similarity regions.

Table 13: Per-task VSI-Bench results across VLM backbones. "Feature. Dist." denotes feature distillation; "MVRD" is our relational distillation. "+ DP" marks Dual Pathway. Best per column within each backbone block is in **bold**.

Backbone	Method	Avg.	Abs. Dist.	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Obj. Count	Obj. Size	Appr. Order
			Geometric					**Linguistic Grounding**		
LLaVA-Video-7B	SFT	57.7	43.6	63.7	64.9	68.9	40.7	70.6	**70.8**	**38.5**
	Feature. Dist.	58.5	48.3	68.4	61.4	75.7	**44.9**	71.4	67.8	30.1
	MVRD	59.9	**50.5**	**69.3**	62.7	**80.0**	40.7	**72.9**	68.3	35.1
	MVRD + DP	**60.4**	49.6	68.6	64.4	78.2	42.8	71.8	70.1	37.7
InternVL3-8B	SFT	50.3	41.8	52.9	57.5	53.2	25.8	70.8	67.5	33.3
	Feature Dist.	57.0	48.2	60.9	62.0	77.1	37.6	70.8	67.2	32.4
	MVRD	58.7	48.7	**63.4**	61.8	**80.1**	**47.4**	**71.8**	68.7	27.4
	MVRD + DP	**59.9**	**50.3**	**63.4**	**63.5**	76.8	44.3	71.7	**68.9**	**40.1**
Qwen2.5-VL-7B	SFT	56.1	47.0	53.7	61.8	73.8	**45.9**	70.5	**70.7**	25.1
	Feature Dist.	55.5	45.7	56.9	65.6	72.4	43.3	71.1	66.5	22.5
	MVRD	56.1	48.0	53.1	**66.3**	75.0	40.2	**71.8**	68.5	25.9
	MVRD + DP	**59.8**	**50.9**	**62.2**	63.2	**76.2**	44.3	71.7	68.9	**40.1**

One notable exception is *Appr. Order* on InternVL3, where MVRD incurs a larger performance drop. We hypothesize that this backbone-specific behavior is related to InternVL3's video input format, which serializes frames as `"Frame-1: <image>, Frame-2: <image>,..."`. Because appearance-order reasoning depends on associating visual content with these frame indices, enforcing cross-frame visual relations may partially interfere with this temporal association.

Importantly, our Dual Pathway variant, which separates the parameters used for relational learning, mitigates this degradation and provides more consistent improvements across tasks, with the largest gain on Qwen2.5-VL-7B (+**3.7%** accuracy).

Ablation on multi-view relation learning. We evaluate two design choices in Section 4: (i) the choice of *relation construction* - how the pairwise relation matrix is computed in each feature space; and (ii) the choice of *matrix loss* - how the student and teacher relation matrices are aligned. Concretely, given features $\{f_i\}_{i=1}^{N}$ from either the VLM or the geometric teacher, we consider two ways of forming the $N \times N$ relation matrix R:

- **Cosine similarity** (used in MVRD): $R_{ij} = \cos(f_i, f_j)$, computed for both the ℓ_2-normalized student and teacher features.
- **Euclidean distance**: $R_{ij} = \|f_i - f_j\|_2$, computed on the raw feature vectors in each space.

For aligning the student matrix R^V with the teacher matrix R^G, we consider:

- **Row-wise Pearson correlation** (used in MVRD), computed between corresponding rows of the student and teacher relation matrices (Appendix B.3).
- **Mean squared error (MSE)**, computed entry-wise as $\frac{1}{N^2}\sum_{i,j}(R^V_{ij} - R^G_{ij})^2$, which directly matches the absolute relation values.

The results are summarized in Table 14. Other components - distillation layer, LoRA configuration, and frame sampling - are kept the same as in the visual-spatial reasoning setup described in Appendix D.1.

Observations. Row-wise Pearson outperforms MSE under both relation constructions, with a much larger gap for Euclidean distance ($+10.5$) than for cosine similarity ($+1.5$). Euclidean distances depend on feature magnitudes, which may differ substantially between the student and teacher. MSE penalizes these scale differences even when their relational patterns agree, whereas row-wise Pearson

Table 14: Ablation on the construction of the cross-view relation matrix and the distillation loss. VSI-Bench average accuracy (↑). The MVRD configuration is in **bold**.

	Row-wise Pearson	MSE
Pair-wise cosine similarity	**59.9**	58.4
Pair-wise Euclidean distance	59.1	48.6

correlation (used in MVRD) is invariant to row-wise shifts and positive rescaling, allowing differences in scale and offset between student and teacher relation values. This may explain the larger gain of row-wise Pearson correlation over MSE for Euclidean distance, since cosine similarity already normalizes feature magnitudes.

Likewise, cosine similarity outperforms Euclidean distance under both losses, suggesting that removing feature-magnitude differences makes relations easier to distill between the VLM and teacher. The worst-performing combination (Euclidean + MSE, 48.6) falls well below the SFT baseline of 57.7. Hence, this ablation indicates that both cosine similarity and Pearson correlation improve accuracy.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We conduct intensive experiments across different models under a fair training setting and different benchmarks to make our claim.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We provide our discussion on limitation and furture work in the appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [No]

Justification: we do not claim any theoretical result.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: we provide full details about the training setting, including base models, hyperparameters, dataset we used for training, which are all publicly available.

   Guidelines:

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

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

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

Answer: [No]

Justification: We have a plan to open-source code after the submission period. Our codebase is not ready to open-source yet.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: we provide all main details in the main text, and all information in the appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: we provide experiment results following the conventional approach in the research field, which does not include the error bars.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: we provide information about GPU, batch size, training epoch, and training hours.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: we conform in every respect, with the NeurIPS Code of Ethics

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: our work is conducted for tasks under an academic setting, which has no direct societal impacts.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We do not see immediate risk of using our model.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite relevant prior work

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: we do not introduce new assets

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: we do not do any crowd sourcing nor research with human subjects

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: we do not do any crowd sourcing nor research with human subjects

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: We only use LLM for writing, editing, formatting, and coding assistance.

Guidelines:

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