3DRS: MLLMs Need 3D-Aware Representation Supervision for Scene Understanding

Xiaohu Huang[1] **Jingjing Wu**[2] **Qunyi Xie**[2] **Kai Han**[1]*
[1] Visual AI Lab, The University of Hong Kong
[2] Department of Computer Vision Technology (VIS), Baidu Inc.
huangxiaohu@connect.hku.hk, kaihanx@hku.hk

Abstract

Recent advances in scene understanding have leveraged multimodal large language models (MLLMs) for 3D reasoning by capitalizing on their strong 2D pretraining. However, the lack of explicit 3D data during MLLM pretraining limits 3D representation capability. In this paper, we investigate the 3D-awareness of MLLMs by evaluating multi-view correspondence and reveal a strong positive correlation between the quality of 3D-aware representation and downstream task performance. Motivated by this, we propose 3DRS, a framework that enhances MLLM 3D Representation learning by introducing Supervision from pretrained 3D foundation models. Our approach aligns MLLM visual features with rich 3D knowledge distilled from 3D models, effectively improving scene understanding. Extensive experiments across multiple benchmarks and MLLMs—including visual grounding, captioning, and question answering—demonstrate consistent performance gains. Project page: https://visual-ai.github.io/3drs

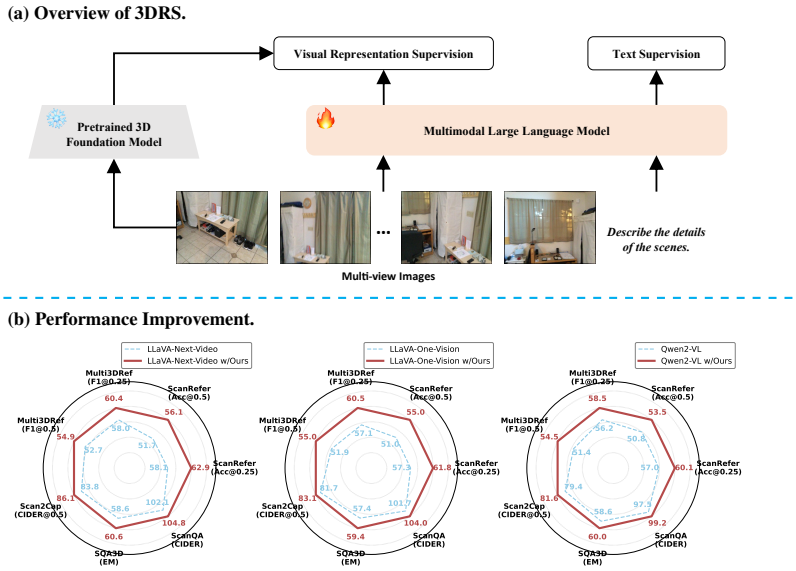


Figure 1: **Enhancing 3D awareness of MLLMs to improve downstream performance.** (a) Besides the common text supervision for MLLMs, 3DRS adopts 3D foundation models to supervise 3D-aware visual representation learning in MLLMs. (b) Combined with 3DRS, we achieve consistent performance improvement across multiple MLLMs and benchmarks.

*Corresponding author.

Preprint. Under review.

1 Introduction

Scene understanding serves as a cornerstone for interpreting 3D environments, enabling a wide range of critical applications ranging from robotic navigation to augmented reality. The recent emergence of large language models (LLMs) [42, 17, 18] has sparked innovative research aimed at endowing these models with scene comprehension capabilities. One major line of research [50, 24, 36, 19, 24, 40, 41, 9, 14, 8, 54] utilizes point cloud encoders—either independently or in combination with multi-view images—to extract 3D representations, which are subsequently projected into a language-aligned space for LLMs. However, these approaches are constrained by the scarcity of paired 3D-text datasets, which impedes effective cross-modal feature alignment.

In response to this challenge, recent state-of-the-art methods [64, 62, 16, 22, 37] have shifted towards leveraging multi-view images exclusively, drawing inspiration from the success of large-scale visual-language pretraining in multimodal LLMs (MLLMs) [29, 27, 59, 1, 46**?**]. These approaches aim to transfer 2D visual understanding to 3D scene comprehension by injecting 3D priors, such as 3D positional embeddings, into the models, thereby allowing MLLMs to capitalize on their extensive pretrained 2D knowledge for 3D interpretation.

Despite these advancements, genuine 3D scene understanding fundamentally requires models to capture intrinsic 3D attributes and spatial structures to comprehend scenes. The absence of explicit 3D data during MLLM pretraining reveals a significant gap, which motivates our core investigation centered around the following questions: (1) *How can we evaluate the ability of MLLMs to learn 3D-aware representations?* (2) *How does the quality of 3D feature learning influence downstream scene understanding performance?* (3) *What methods can enhance 3D-aware representation learning within MLLM frameworks?* While several prior works [53, 28, 15, 33] have attempted to probe the 3D awareness of 2D vision foundation models, systematic investigation into 3D-aware representation learning in MLLMs remains largely unexplored. This gap is particularly crucial given the growing adoption of MLLMs in multimodal 3D understanding tasks. Our study aims to address this overlooked area and provide new insights into 3D representation learning within the MLLM paradigm.

For the first question, we conduct comprehensive experiments to evaluate the 3D awareness on three representative MLLMs, including LLaVA-Next-Video [59], LLaVA-One-Vision [27], and Qwen2-VL [46], following the finetuing settings of Video-3D LLM [62]. Specifically, we assess 3D awareness via view equivariance, quantifying it by computing the feature similarity between corresponding pairs from the same 3D voxel across different views. This evaluation requires MLLMs to associate the same object across multiple views, thereby reflecting their capacity for 3D representation. Our analysis encompasses six datasets spanning tasks such as 3D grounding [5], captioning [12], and question answering [2].

To address the second question, we systematically analyze model performance across these datasets and observe that *samples with higher correspondence scores—i.e., those exhibiting stronger 3D awareness—consistently lead to improved performance.* This finding demonstrates a strong positive correlation between the quality of 3D-aware representations and downstream scene understanding performance, highlighting the necessity of enhancing 3D feature learning in MLLMs.

In response to the third question and building upon our earlier findings, we first introduce a view equivalence supervision strategy for MLLMs, encouraging alignment between feature pairs corresponding to the same 3D location across different views (positive pairs) while discouraging similarity among unrelated pairs (negative pairs). While this approach results in some performance gains, the supervision provided by such handcrafted, single-task objectives is inherently limited for 3D learning.

In contrast, recent 3D foundation models such as VGGT [45] and FLARE [57] are pretrained end-to-end on multi-view image sequences spanning a diverse set of 3D geometric tasks—including not only correspondence learning, but also depth estimation and camera parameter prediction. This comprehensive pretraining enables them to encode rich 3D properties within their features. Building on this, we propose a framework, 3DRS, that leverages these pretrained models by using their features as alignment targets for the visual outputs of MLLMs, thereby facilitating more effective 3D-aware representation learning. Unlike previous 3D MLLM approaches, in addition to traditional text token supervision, our framework employs explicit 3D-specific supervision directly on scene visual tokens. As demonstrated in our experiments (see Fig. 1), incorporating this form of 3D supervision consistently improves performance across a range of MLLMs and benchmarks. Notably, our approach incurs no additional training overhead, since the supervisory features can be pre-extracted offline.

We believe this design offers valuable new insights for applying 3D foundation models in scene understanding. The key contribution of this paper can be summarized as follows:

- We conduct a systematic evaluation of the 3D-awareness of MLLMs using multi-view correspondence metrics, and observe a strong positive correlation between 3D-aware representation quality and downstream scene understanding performance across diverse tasks, datasets, and models.
- We propose a 3D-aware representation supervision framework that aligns MLLM visual features with those of a 3D geometry-pretrained model, enabling effective 3D feature learning.
- Extensive experiments demonstrate consistent performance improvements across multiple MLLMs and 3D scene understanding benchmarks, validating the effectiveness and generality of our approach.

2 Method

2.1 Investigating 3D-Aware Representation Learning in MLLMs

2.1.1 Preliminaries

A MLLM typically consists of two main components: an image encoder $\mathcal{E}_{\text{img}}$ and a text decoder $\mathcal{T}$. In this work, the input to our MLLM comprises a set of N multi-view images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, each associated with per-pixel 3D coordinates $\mathcal{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_N\}$, where $\mathbf{C}_i \in \mathbb{R}^{H\times W\times 3}$ for image I_i of size $H \times W$. The 3D coordinates for each pixel are computed from the depth map and the corresponding camera intrinsic and extrinsic parameters; detailed formulas and procedures can be found in the App. A.1.

The MLLM receives both multi-view images and language instructions as input. Internally, for each image I_i, the image encoder $\mathcal{E}_{\text{img}}$ extracts visual features $\mathbf{F}_i \in \mathbb{R}^{H\times W\times d}$, where d is the feature dimension. Following Video3DLLM [62], we encode the per-pixel 3D coordinates via a positional encoding function $\phi(\cdot)$ and inject this information into the image features by addition:

$$\mathbf{F}_i^{3D} = \mathbf{F}_i + \phi(\mathbf{C}_i).$$

This design allows the MLLM to inherit 2D perceptual knowledge from pretraining while equipping it with explicit 3D priors.

During finetuning, the MLLM—which we denote as f_θ—passes visual features $\{\mathbf{F}_i^{3D}\}_{i=1}^N$ with the instruction tokens to the text decoder for autoregressive text generation. After the processing of the text decoder, we refer to the final per-pixel visual embedding of pixel p in image I_i from LLM as $\mathbf{f}_i(p)$. The model is optimized by minimizing the standard cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, \{I_i, \mathbf{C}_i\}_{i=1}^N, \text{instruction}),$$

where y_t is the t-th output token, and p_θ is the probability predicted by the model given all previous tokens and the multimodal context (*i.e.*, images and instructions).

2.1.2 Assessing 3D Feature Learning via Multi-View Correspondence

Inspired by the crucial role of cross-view correspondences in 3D modeling [20], we propose a correspondence-based evaluation framework. Multi-view correspondences are fundamental in 3D vision, serving as essential cues for core tasks such as ray retriangulation [20], bundle adjustment [43], and pose estimation [39]. They are also critical for downstream applications like instance recognition and retrieval [39, 51, 49]. Therefore, we adopt multi-view correspondence analysis as a proxy to evaluate the 3D representations of MLLMs. This approach requires the model to accurately associate and align objects or regions that occupy the same position in 3D space across different viewpoints.

Voxelization and correspondence pair construction. We first voxelize the 3D scene into a regular grid of voxels $\mathcal{V} = \{v_1, \ldots, v_M\}$. For each view I_i, given its per-pixel 3D coordinates $\mathbf{C}_i$, we assign

every pixel's feature $\mathbf{f}_i(p)$ to a voxel according to its 3D position. Features from different views that fall into the same voxel v_k are regarded as *correspondence pairs*.

Feature similarity and correspondence scores. Let $\mathcal{P}_k$ denote all correspondence feature pairs in voxel v_k, *i.e.*, all pairs $(\mathbf{f}_i(p), \mathbf{f}_j(q))$ where both pixels p and q from images I_i and I_j are assigned to v_k with $i \neq j$. For any pair of visual features $(\mathbf{f}_a, \mathbf{f}_b)$ from the last layer of MLLM, feature similarity is measured by the cosine similarity:

$$S(\mathbf{f}_a, \mathbf{f}_b) = \frac{\mathbf{f}_a^\top \mathbf{f}_b}{\|\mathbf{f}_a\| \cdot \|\mathbf{f}_b\|}.$$

For each sequence, we compute:

$$\bar{S} = \frac{1}{|\mathcal{P}|} \sum_{(\mathbf{f}_a, \mathbf{f}_b) \in \mathcal{P}} S(\mathbf{f}_a, \mathbf{f}_b),$$

where $\bar{S}$ and $\mathcal{P}$ denote the *correspondence score* for each sequence and all the correspondence pairs in this sequence. A higher correspondence score indicates that the model produces more consistent features across views for the same 3D spatial location, reflecting stronger 3D-aware representation learning.

2.1.3 Quality of 3D Feature vs. Downstream Task Performance.

We evaluate three representative MLLMs, LLaVA-Next-Video, LLaVA-OneVision, and Qwen2-VL, on five diverse 3D scene understanding benchmarks, including visual grounding (Multi3DRefer, ScanRefer), captioning (Scan2Cap), and question answering (ScanQA, SQA3D). All benchmarks are based on multi-view RGBD sequences. The three MLLMs respectively emphasize video understanding, joint image-video reasoning, and advanced arbitrary-resolution visual encoding.

To analyze the relationship between 3D feature learning and downstream task performance, we sort samples within each dataset by their correspondence scores and divide them into four quartiles (Q1–Q4, lowest to highest). From Fig. 2, we observe a clear trend: *as the correspondence score increases, the model's performance on the downstream task consistently improves.* This strong positive correlation demonstrates the critical importance of 3D-aware representation quality for effective scene understanding in MLLMs.

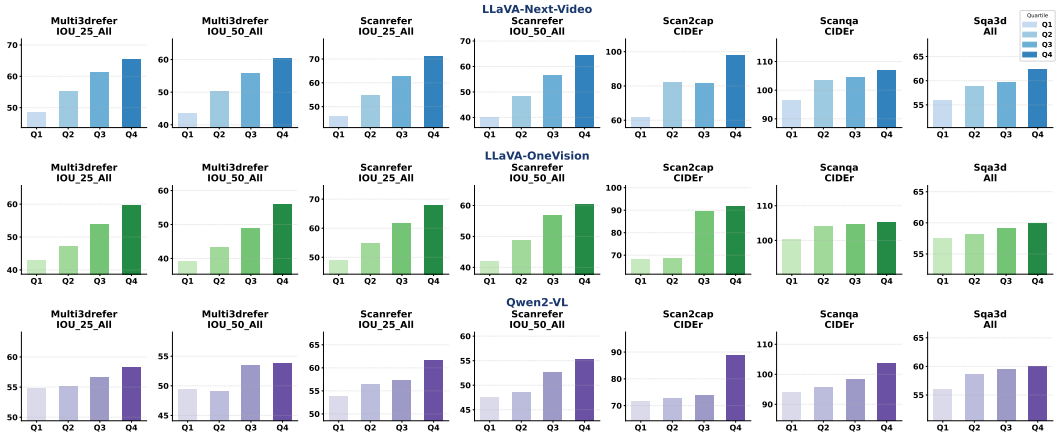


Figure 2: **Performance across correspondence score quartiles.** Model performance across correspondence score quartiles (Q1–Q4, lowest to highest) for each dataset. Samples were divided into quartiles by their correspondence scores. A clear trend is observed: model accuracy improves as the correspondence score increases.

These findings highlight the need for strategies to further enhance 3D-aware representation learning in MLLMs, which we address in the next section.

2.2 Enhancing 3D-Aware Representation Learning in MLLMs

2.2.1 Correspondence-based 3D Supervision Loss

Inspired by our correspondence-based evaluation, a straightforward approach is to directly supervise the MLLM's visual features to be consistent for matched 3D locations across views and dissimilar for mismatched locations. We let $\mathcal{P}_k^+$ denote all positive feature pairs in voxel v_k, *i.e.*, all pairs $(\mathbf{f}_i(p), \mathbf{f}_j(q))$ where pixels p and q from images I_i and I_j are assigned to v_k with $i \neq j$. Similarly, $\mathcal{P}_k^-$ denotes negative pairs between v_k and any other voxel v_l ($l \neq k$). We supervise these objectives directly using a simple loss function by maximizing the feature similarity in $\mathcal{P}^+$ and minimizing that in $\mathcal{P}^-$:

$$\mathcal{L}_{\text{corr}}^+ = \frac{1}{|\mathcal{P}^+|} \sum_{(\mathbf{f}_a, \mathbf{f}_b) \in \mathcal{P}^+} \left[1 - S(\mathbf{f}_a, \mathbf{f}_b)\right],$$

$$\mathcal{L}_{\text{corr}}^- = \frac{1}{|\mathcal{P}^-|} \sum_{(\mathbf{f}_a, \mathbf{f}_b) \in \mathcal{P}^-} S(\mathbf{f}_a, \mathbf{f}_b).$$

The overall correspondence loss is a weighted sum:

$$\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{corr}}^+ + \mathcal{L}_{\text{corr}}^-.$$

By directly supervising positive pairs to be similar and negative pairs to be dissimilar, this correspondence loss encourages the model to learn multi-view 3D correspondences, thus enhancing the 3D-awareness of the learned representations. As will be shown in the experiments Sec. 3, supplementing the standard cross-entropy objective with $\mathcal{L}_{\text{corr}}$ leads to improvements in downstream task performance. However, as this loss primarily targets view equivariance, the range of 3D properties captured remains limited, motivating the need for richer supervision.

2.2.2 3D Foundation Model-Guided Feature Distillation

To overcome the inherent limitations of single-task supervision, we further introduce a knowledge distillation framework, 3DRS, that leverages the rich 3D priors embedded in 3D foundation models, e.g, FLARE and VGGT. These models are pretrained on a wide array of 3D geometric tasks—including correspondence learning, camera parameter estimation, multi-view depth prediction, and dense point cloud reconstruction—which enables them to extract robust and highly 3D-aware visual features from multi-view image sequences.

Distillation target preparation. As shown in Fig. 3a, given a set of multi-view images $\mathcal{I}$ for a scene, we first input them into a pretrained 3D foundation model g, which outputs a collection of per-pixel visual features $\{\mathbf{f}_i^{\text{3D}}(p)\}$ for each image I_i and pixel p. Since the spatial resolution of these features may differ from those of the MLLM outputs $\{\mathbf{f}_i(p)\}$, we apply 2D average pooling to the 3D foundation model's output to match the MLLM feature map size.

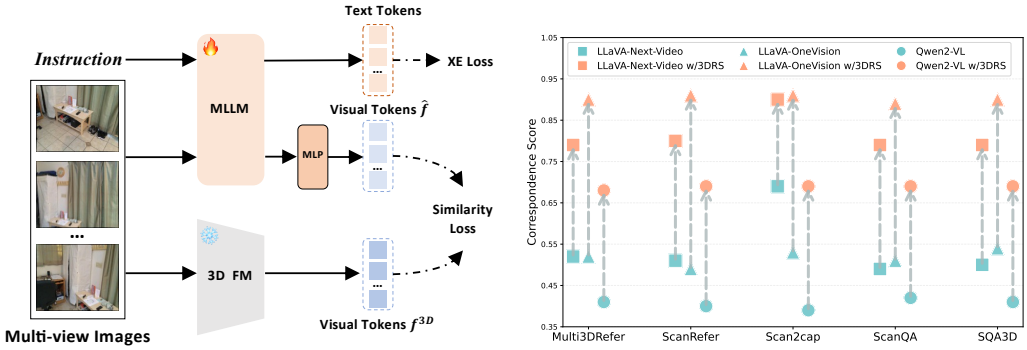
Feature alignment and loss. To align the MLLM's per-pixel visual features with the 3D foundation model, we first process each $\mathbf{f}_i(p)$ with a two-layer MLP (denoted as $\text{MLP}_{\text{align}}$) to ensure compatibility in feature dimension:

$$\hat{\mathbf{f}}_i(p) = \text{MLP}_{\text{align}}(\mathbf{f}_i(p)).$$

We then employ a distillation loss based on euclidean similarity to maximize the alignment between the MLLM features $\hat{\mathbf{f}}_i(p)$ and the corresponding 3D foundation model features $\mathbf{f}_i^{\text{3D}}(p)$:

$$\mathcal{L}_{\text{align}} = -\frac{1}{NHW} \sum_{i=1}^{N} \sum_{p \in I_i} S\left(\hat{\mathbf{f}}_i(p), \mathbf{f}_i^{\text{3D}}(p)\right),$$

where $S(\cdot, \cdot)$ denotes cosine similarity, and the sum is calculated over all pixels and views in the batch.



(a) **Details of 3DRS.** (b) **Comparison of correspondence learning.**

Figure 3: (a) 3DRS uses a 3D foundation model to supervise the visual representation of the MLLM. (b) 3DRS effectively improves the correspondence learning for MLLMs.

Overall training objective. The final training objective for the MLLM combines the standard cross-entropy loss for text generation and the 3D foundation model distillation loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{align}}.$$

This approach enables the MLLM to inherit comprehensive 3D knowledge from powerful geometry-pretrained models, facilitating the learning of richer and more robust 3D-aware representations. Importantly, the distillation targets from the 3D foundation model can be precomputed offline, introducing no additional overhead during MLLM fine-tuning.

As illustrated in Fig. 3b, we compare the correspondence scores before and after applying our 3DRS, where VGGT serves as the foundation model. The results consistently demonstrate that introducing 3DRS leads to substantial improvements in correspondence learning ability across all evaluated MLLMs and benchmarks. This proves the effectiveness of leveraging a pretrained 3D foundation model as a teacher model for enhancing 3D-aware representation learning in MLLMs. More comprehensive experimental results and analyses are detailed in Sec. 3.

3 Experiments

3.1 Datasets and Evaluation Metrics

Datasets. We evaluate our approach on six benchmarks that collectively span key challenges in 3D scene understanding. ScanRefer [5] focuses on localizing objects using free-form language, while Multi3DRefer [58] generalizes this to queries referencing zero, one, or multiple objects, better reflecting real-world ambiguity. Scan2Cap [12] addresses dense captioning by pairing detected objects in 3D scans with natural language descriptions. For question answering, ScanQA [2] tasks models with answering open-ended questions grounded in 3D geometry and semantics, and SQA3D [32] goes further by requiring situated reasoning: agents must interpret their position and context to answer complex queries. All these datasets are sourced from the richly annotated ScanNet [13] corpus, and we follow standard validation and test splits as established in prior work [24, 64, 9, 62]. Besides, VSI-Bench [52] is used to evaluate the performance on visual-based spatial understanding tasks, which are composed of numerical and multiple-choice questions. The statistics of training sets are detailed in the App. A.2.

Evaluation metrics. For ScanRefer, we report accuracy at IoU thresholds of 0.25 and 0.5 (Acc@0.25, Acc@0.5). Multi3DRefer uses F1 scores at matching IoU thresholds. Scan2Cap is evaluated by CIDEr and BLEU-4 scores at 0.5 IoU (C@0.5, B-4@0.5). ScanQA is assessed by CIDEr and exact match accuracy (C, EM), while SQA3D uses exact match accuracy as the metric.

3.2 Implementation Details

Our experiments leverage several MLLMs, including LLaVA-Next-Video 7B [59], LLaVA-OneVision 7B [27], and Qwen2-VL 7B [46]. In addition to these baselines, we systematically compare the

Table 1: Performance comparison on 3D scene understanding benchmarks. Specialists are single-task methods, while generalists target multiple tasks. Bold denotes best performance.

Method	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
Specialists									
ScanRefer [5]	37.3	24.3	–	–	–	–	–	–	–
MVT [26]	40.8	33.3	–	–	–	–	–	–	–
3DVG-Trans [60]	45.9	34.5	–	–	–	–	–	–	–
ViL3DRel [7]	47.9	37.7	–	–	–	–	–	–	–
M3DRef-CLIP [58]	51.9	44.7	42.8	–	38.4	–	–	–	–
Scan2Cap [12]	–	–	–	–	35.2	22.4	–	–	–
ScanQA [2]	–	–	–	–	–	–	64.9	21.1	47.2
3D-VisTA [65]	50.6	45.8	–	–	66.9	34.0	69.6	22.4	48.5
Generalists									
3D-LLM(Flamingo) [22]	21.2	–	–	–	–	–	59.2	20.4	–
3D-LLM(BLIP2-flant5) [22]	30.3	–	–	–	–	–	69.4	20.5	–
Chat-3D [47]	–	–	–	–	–	–	53.2	–	–
Chat-3D v2 [24]	42.5	38.4	45.1	41.6	63.9	31.8	87.6	–	54.7
LL3DA [8]	–	–	–	–	62.9	36.0	76.8	–	–
SceneLLM [16]	–	–	–	–	–	–	80.0	27.2	53.6
LEO [25]	–	–	–	–	72.4	38.2	101.4	21.5	50.0
Grounded 3D-LLM [9]	47.9	44.1	45.2	40.6	70.6	35.5	72.7	–	–
PQ3D [66]	57.0	51.2	–	50.1	80.3	36.0	–	–	47.1
ChatScene [24]	55.5	50.2	57.1	52.4	77.1	36.3	87.7	21.6	54.6
LLaVA-3D [64]	54.1	42.4	–	–	79.2	41.1	91.7	27.0	55.6
Inst3D-LLM [54]	57.8	51.6	58.3	53.5	79.7	38.3	88.6	24.6	–
3D-LLaVA [14]	51.2	40.6	–	–	78.8	36.9	92.6	–	54.5
Video-3D LLM [62]	58.1	51.7	58.0	52.7	83.8	41.3	102.1	30.1	58.6
3DRS	**62.9**	**56.1**	**60.4**	**54.9**	**86.1**	**41.6**	**104.8**	**30.3**	**60.6**

effect of using 2D versus 3D foundation models as teachers for MLLM finetuning. The 2D teacher models include DINOv2 [34], MAE [21], and SigLIP [44], while the 3D teacher models comprise FLARE [57] and VGGT [45]. Unless stated otherwise, we use LLaVA-Next-Video as the MLLM and VGGT as the representation teacher for our experiments.

For both training and inference, we uniformly sample 32 frames per scan to construct multi-view image sets. For evaluating the correspondence score, we use the voxel size of 0.1 for voxelization. All models are optimized using Adam, with a batch size of 16 and a warm-up ratio of 0.03. The learning rates are set to a maximum of 1×10^{-5} for the language model and 2×10^{-6} for the visual backbone during the warm-up period. During training for visual grounding and dense captioning, ground truth object regions are used as candidates, whereas during inference, we follow the procedure of [24, 25, 62] and employ Mask3D [38] to generate object proposals. For LLaVA-Next-Video and LLaVA-OneVision, we finetune all model parameters. For Qwen2-VL, due to GPU memory constraints, we finetune only the projector and the LLM components. We use 8 H100 NVIDIA GPUs for all experiments.

3.3 Comparison with State-of-the-Art Models

Table 1 presents a comprehensive comparison between our approach, task-specific specialist models—which require fine-tuning on individual datasets—and 3D generalist models that are capable of handling multiple tasks. Compared to specialist models, our approach achieves substantial performance improvements. This demonstrates the significant benefits brought by joint training and the LLM-based architecture, which contribute to superior generalization and feature integration compared to methods tailored for specific tasks. Furthermore, our method consistently outperforms 3D generalist approaches that utilize point clouds as input, such as LL3DA, Chat-3D, Grounded 3D-LLM, and 3D-LLaVA. Compared to Inst3D-LLM—which fuses multi-view images and point clouds—our approach also shows clear advantages, highlighting the strength of leveraging MLLMs as the backbone. Additionally, our method achieves considerable improvements over other MLLM-based methods, including LLaVA-3D and Video-3D LLM. These results collectively indicate that enhancing the 3D-awareness of MLLMs is highly effective for 3D scene understanding tasks, further validating the effectiveness of our proposed strategy.

Table 2: Performance comparison on VSI-Bench.

Method	Avg.	Obj. Count	Abs. Distance	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
GPT-4o [1]	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Pro [?]	45.4	56.2	30.9	**64.1**	43.6	**51.3**	**46.3**	**36.0**	34.6
LongVA-7B [56]	29.2	38.0	22.2	33.1	43.3	25.4	15.7	33.1	17.7
InternVL2-40B [11]	37.0	41.3	26.2	48.2	27.5	47.6	32.7	27.8	44.7
LLaVA-Video-7B [59]	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-Video-72B [59]	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
LLaVA-OneVision-7B [27]	32.4	47.7	20.7	47.4	42.5	35.2	29.4	24.4	
LLaVA-OneVision-72B [27]	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
3DRS	**45.9**	**68.7**	**34.8**	53.6	**56.6**	40.9	43.2	30.4	**39.2**

Table 2 compares our method with leading proprietary APIs and open-source models on VSI-Bench [52], covering tasks such as object counting, spatial distance estimation, size reasoning, and sequence understanding. 3DRS achieves the best open-source results on most metrics—including object count, absolute distance, room size, and appearance oder—and remains competitive with proprietary models. These results demonstrate the strong spatial reasoning, generalization, and comprehensive scene understanding capabilities of our approach across diverse 3D vision tasks.

Table 3: Performance comparison of 3DRS when using with different MLLMs.

Method	ScanRefer		Multi3DRef		Scan2Cap		ScanQA		SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	C	EM	EM
LLaVA-Next-Video [59]	58.1	51.7	58.0	52.7	41.3	83.8	102.1	30.1	58.6
LLaVA-Next-Video w/ 3DRS	**62.9**	**56.1**	**60.4**	**54.9**	**41.6**	**86.1**	**104.8**	**30.3**	**60.6**
LLaVA-OneVision [27]	57.3	51.0	57.1	51.9	40.4	81.7	101.7	29.0	57.4
LLaVA-OneVision w/ 3DRS	**61.8**	**55.0**	**60.5**	**55.0**	**41.2**	**83.1**	**104.0**	**29.5**	**59.4**
Qwen2-VL [46]	57.0	50.8	56.2	51.4	39.5	79.4	97.5	28.7	58.6
Qwen2-VL w/ 3DRS	**60.1**	**53.5**	**58.5**	**54.5**	**40.9**	**81.6**	**99.2**	**28.9**	**60.0**

3.4 Diagnostic Study

Effectiveness with different MLLMs. Table 3 demonstrates that integrating 3DRS with different MLLMs—LLaVA-Next-Video, LLaVA-OneVision, and Qwen2-VL—consistently boosts performance across all evaluated benchmarks. For example, LLaVA-Next-Video w/ 3DRS improves ScanRefer Acc@0.25 from 58.1 to 62.9, and Multi3DRef F1@0.25 from 58.0 to 60.4. Similar gains are observed for LLaVA-OneVision and Qwen2-VL, where 3DRS brings improvements on every dataset and metric. These results highlight the general applicability of our approach and its effectiveness in enhancing 3D scene understanding for various MLLMs.

Comparison between 2D and 3D foundation models. Table 4 compares the performance of using 2D and 3D foundation models as representation supervisors. It is clear that 3D foundation models (FLARE and VGGT) outperform all 2D foundation models (MAE, Siglip2, Dinov2) across almost every metric. This performance gap can be attributed to the inherent difference in the prior knowledge captured by 2D and 3D foundation models. 3D models are pre-trained on large-scale 3D data and thus better capture geometric structure, spatial relationships, and depth information, which are critical for 3D scene understanding tasks. In contrast, 2D foundation models, trained on images, lack explicit 3D spatial priors and struggle to provide effective supervision for learning 3D-aware representations. This highlights the importance of 3D-specific foundation models for achieving strong results in downstream 3D tasks.

Comparison of supervision signal. Table 5 shows that using correspondence loss for supervision brings improvements over the baseline, demonstrating the effectiveness of encouraging the model to learn multi-view correspondences. However, when 3D foundation model supervision is applied, the performance increases even further across all metrics. This indicates that 3D foundation models, with their rich 3D prior knowledge learned during pre-training, can more effectively enhance the 3D representation ability of MLLMs and yield greater gains for 3D understanding tasks.

Comparison of supervision at different layers. Table 6 examines the effect of applying 3D foundation model supervision at different layers of the network. The results reveal that supervision

at deeper layers, especially the last layer, leads to the highest performance. This is likely because deeper layers are closer to the output and thus have a more direct impact on the final predictions. Additionally, these layers possess more parameters and a greater capacity to fit 3D features, which results in larger improvements on downstream tasks.

Table 4: Ablation study on using different 2D/3D foundation models as the representation supervisor. Bold denotes the best in each group.

Representation Supervisor	ScanRefer		Multi3DRef		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
Baseline	58.1	51.7	58.0	52.7	83.8	41.3	102.1	30.1	58.6
2D Foundation Models									
Siglip2 [44]	58.2	52.9	59.7	53.1	81.7	40.2	100.2	29.1	59.4
MAE [21]	59.1	**53.7**	**60.0**	**53.7**	**82.8**	**40.4**	102.5	29.5	59.2
Dinov2 [34]	**59.8**	53.3	58.5	53.5	80.3	39.3	**103.5**	**29.6**	**60.1**
3D Foundation Models									
FLARE [57]	62.1	55.7	59.8	54.8	**86.6**	**42.5**	104.4	30.1	60.1
VGGT [45]	**62.9**	**56.1**	**60.4**	**54.9**	86.1	41.6	**104.8**	**30.3**	**60.6**

4 Related Work

4.1 Scene Understanding with Large Language Models

LLMs, owing to their strong reasoning capabilities and remarkable success in 2D image understanding, have been widely applied to scene understanding tasks. Early works such as PointLLM [50], Point-Bind [19], GPT4Point [36], MiniGPT-3D [40], and Chat-3D [47] leverage the alignment between point cloud and text features to facilitate 3D scene comprehension. Building on this foundation, methods like Grounded 3D-LLM [9], LL3DA [8], 3D-LLaVA [14], and Inst3D-LLM [54] design more advanced cross-modal modules to better fuse multi-modal features, thereby enhancing scene representations. Furthermore, Chat-Scene [24] and Inst3D-LLM [54] exploit the complementary nature of 2D and 3D features to further boost scene understanding.

Some recent approaches, such as 3D-LLM [22] and Scene LLM [16], employ multi-view inputs and introduce 3D priors to transform 2D representations into a 3D-aware format. Thanks to pre-training on large-scale image-text datasets, methods based on MLLMs are gaining increasing popularity in the field of scene understanding. For instance, LLaVA-3D [64] takes multi-view images as input and utilizes voxelization to reduce the dimensionality of representations, thus lowering computational costs while leveraging the strengths of MLLMs. However, many MLLMs require specially structured inputs, making them incompatible with certain approaches. Video 3D-LLM [62] and GPT4Scene [37] more naturally inherit the MLLM pipeline by introducing 3D priors—such as positional embeddings or spatial markers—enabling the model to better comprehend 3D scene content.

Our work follows this line of MLLM-based scene understanding, aiming to probe the 3D-awareness of MLLMs and analyze their relationships with downstream tasks. In particular, we demonstrate that introducing guidance from 3D foundation models can effectively enhance the representational capability of MLLMs for 3D scene understanding.

4.2 3D-Awareness in Vision Models

Several studies have investigated 3D-awareness; however, most prior work has focused on pure vision models rather than MLLMs, and primarily leveraged 2D foundation models instead of 3D ones. For example, FiT3D [55], Probe3D [15], and Lexicon3D [33] empirically analyze the 3D-awareness of visual foundation models. CUA-O3D [28] proposes integrating multiple 2D foundation models for 3D scene understanding, while Yang et al.[53] evaluates and enhances the 3D-awareness of ViT-based models in various downstream tasks. Some previous 3D detection works [10, 48, 3] have focused on improving 3D representations for pure vision models or CLIP-style vision-language models (VLMs), primarily aiming to enhance geometric understanding and spatial localization within unimodal or early multimodal frameworks. In addition, several studies on scene understanding [30, 23, 35] have

Table 5: Comparison of different supervision strategies.

Method	Multi3DRef		ScanRefer	
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
Baseline	58.0	52.7	58.1	51.7
w/ Correspondence Loss	60.1	53.3	59.1	53.7
w/ 3D Supervision	**62.9**	**56.1**	**60.4**	**54.9**

Table 6: 3D foundation model supervision at different layers.

Layer	Multi3DRef		ScanRefer	
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
Last Layer	**62.9**	**56.1**	**60.4**	**54.9**
3rd Last Layer	61.7	54.9	59.7	54.3
5th Last Layer	61.4	54.8	59.3	54.0
10th Last Layer	59.1	53.6	53.3	53.8

investigated various strategies for distilling 3D representations, such as transferring knowledge from 3D models to 2D networks or promoting cross-modal alignment. However, these efforts have not addressed the unique challenges presented by MLLMs, which require a more holistic integration of visual, linguistic, and spatial information. As a result, the potential of distilling 3D awareness into MLLMs for richer and more generalizable scene understanding remains largely unexplored.

In contrast, our work specifically targets the 3D-awareness of MLLMs. Rather than enhancing 3D feature learning via 2D foundation models, we introduce 3D foundation models as supervisors to directly guide and improve the 3D representation capabilities of MLLMs.

5 Conclusion

In this paper, we present a comprehensive study of the 3D representation capabilities of multi-modal large language models (MLLMs) in the context of scene understanding. While most existing research has centered on leveraging 2D foundation models to improve visual reasoning in MLLMs, the role and utility of 3D foundation models in this setting remain largely unexplored. To bridge this gap, we propose 3DRS, a novel framework that introduces direct 3D-aware supervision to MLLMs by leveraging pretrained 3D foundation models as teachers. Our approach enables MLLMs to acquire richer geometric and spatial representations, facilitating more accurate and robust understanding of complex 3D scenes. Through extensive experiments on diverse 3D scene understanding benchmarks, we demonstrate that 3DRS consistently improves performance across a variety of tasks, such as object localization, spatial reasoning, and 3D question answering. These results highlight the unique advantages and significant potential of integrating 3D foundation models for advancing multimodal scene understanding.

6 Limitation

While our paper aims to enhance the 3D-awareness of MLLMs, the relatively limited size of the dataset used for finetuning—especially when compared to that used during the MLLM pretraining stage—may restrict the full realization of our approach's potential. Consequently, the improvements demonstrated in this work may only represent an initial step toward more robust 3D understanding. A promising direction for future research is to incorporate 3D-awareness learning into the pretraining stage of MLLMs, which could lead to fundamentally stronger models with deeper 3D comprehension. Besides, due to the distillation-based nature of our approach, the performance of our method is upper-bounded by the quality of the teacher 3D foundation model. Any limitations or failure modes of the teacher—such as inaccurate correspondence, erroneous depth estimation, or incomplete geometric representations—can be propagated to the student MLLM and may potentially mislead it during training. While our experiments demonstrate consistent improvements over strong baselines, it is possible that errors or biases in the teacher's predictions can negatively impact the downstream 3D reasoning abilities of the student model. We believe that, as 3D foundation models continue to rapidly advance, this limitation becomes less pronounced over time.

Acknowledgments

This work is supported by Hong Kong Research Grant Council - General Research Fund (Grant No. 17213825). We would like to thank Weining Ren for the valuable and insightful discussions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. License: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

[3] Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, and Jun Won Choi. Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features. In *CVPR*, 2024.

[4] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, 2022.

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. License: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

[6] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D^3net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *ECCV*, 2022.

[7] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *NeurIPS*, 2022.

[8] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *CVPR*, 2024.

[9] 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.

[10] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022.

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[12] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, 2021. License: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. License: ScanNet Terms of Use.

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

[15] 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 *CVPR*, 2024.

[16] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

[17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

[19] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

[20] Richard Hartley. *Multiple view geometry in computer vision*, volume 665. Cambridge university press, 2003.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[22] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023.

[23] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *ICCV*, 2021.

[24] Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *CoRR*, 2023.

[25] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *ICML*, 2024.

[26] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, 2022.

[27] 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.

[28] Jinlong Li, Cristiano Saltori, Fabio Poiesi, and Nicu Sebe. Cross-modal and uncertainty-aware agglomeration for open-vocabulary 3d scene understanding. In *CVPR*, 2025.

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[30] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3d-to-2d distillation for indoor scene parsing. In *CVPR*, 2021.

[31] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024.

[32] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: situated question answering in 3d scenes. In *ICLR*, 2023. License: CC-BY-4.0.

[33] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. In *NeurIPS*, 2024.

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

[35] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023.

[36] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *CVPR*, 2024.

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

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

[39] Stefan Stojanov, Anh Thai, Zixuan Huang, and James M Rehg. Learning dense object descriptors from multiple views for low-shot category generalization. In *NeurIPS*, 2022.

[40] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *ACM Multi*, 2024.

[41] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Jinfeng Xu, Yixue Hao, Long Hu, and Min Chen. More text, less point: Towards 3d data-efficient point-language understanding. In *AAAI*, 2025.

[42] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(8), 2024.

[43] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*. Springer, 1999.

[44] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[45] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025.

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

[48] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *ICCV*, 2023.

[49] Hao Wu, Ruochong Li, Hao Wang, and Hui Xiong. Com3d: Leveraging cross-view correspondence and cross-modal mining for 3d retrieval. In *ICME*, 2024.

[50] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024.

[51] Yong Xu, Chaoda Zheng, Ruotao Xu, Yuhui Quan, and Haibin Ling. Multi-view 3d shape recognition via correspondence-aware deep learning. *IEEE Transactions on Image Processing*, 2021.

[52] 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 *CVPR*, 2025.

[53] Yang You, Yixin Li, Congyue Deng, Yue Wang, and Leonidas Guibas. Multiview equivariance improves 3d correspondence understanding with minimal feature finetuning. *arXiv preprint arXiv:2411.19458*, 2024.

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

[55] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *ECCV*, 2024.

[56] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[57] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *CVPR*, 2025.

[58] Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*, 2023. License: MIT.

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

[60] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, 2021.

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

[62] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *CVPR*, 2025.

[63] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *CVPR*, 2024.

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

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

[66] 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. In *ECCV*, 2024.

A Technical Appendices and Supplementary Material

A.1 World Coordinate Computation

Given a set of N images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, each image I_i is paired with its depth map $D_i \in \mathbb{R}^{H\times W}$, camera intrinsic matrix $K_i \in \mathbb{R}^{3\times 3}$, and camera-to-world extrinsic matrix $T_i \in \mathbb{R}^{4\times 4}$. For a pixel at (u, v) in image I_i, the corresponding 3D coordinate in the global coordinate system, denoted as $\mathbf{C}_i(u, v) \in \mathbb{R}^3$, is computed as:

$$\begin{bmatrix} \mathbf{C}_i(u,v) \\ 1 \end{bmatrix} = T_i \begin{bmatrix} D_i(u,v) \cdot K_i^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \\ 1 \end{bmatrix} \tag{1}$$

Repeating this process for all pixels yields the per-pixel 3D coordinate map $\mathbf{C}_i \in \mathbb{R}^{H\times W\times 3}$ for each image I_i. The complete set of coordinate maps is denoted as $\mathcal{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_N\}$.

A.2 Datsests for Training

For model fine-tuning, we utilize a collection of well-established 3D vision-language datasets. Specifically, we follow the model finetuning settings of Video-3D LLM [62] by using the validation splits of ScanRefer, Multi3DRefer, Scan2Cap, and ScanQA, as well as the test split of SQA3D. Across these datasets, the number of data samples varies significantly: ScanRefer and Scan2Cap each provide 36,665 samples, while Multi3DRefer offers 43,838 entries. ScanQA contains 26,515 instances, and SQA3D is the largest with 79,445 samples. Most datasets are derived from 562 unique scans, except SQA3D, which includes 518 scans. We further report the average lengths of questions and answers for each dataset. For example, question lengths range from approximately 13 to 38 words, with Scan2Cap and ScanQA also providing answer texts, averaging 17.9 and 2.4 words in length, respectively. In SQA3D, the average question and answer lengths are 37.8 and 1.1 words. For the evaluation on VSI-Bench, we use the pre-training data from VG-LLM [61].

A.3 Detailed Comparison

In this section, we provide a detailed comparison with other methods using all metrics across 5 benchmarks.

Scanrefer. Tab. 7 shows that our method 3DRS achieves the best overall performance on the ScanRefer validation set, especially in the challenging "Multiple" scenario where precise target

Table 7: Performance comparison on the validation set of ScanRefer [5]. "Unique" and "Multiple" depends on whether there are other objects of the same class as the target object.

Method	Unique		Multiple		Overall	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [5]	76.3	53.5	32.7	21.1	41.2	27.4
MVT [26]	77.7	66.4	31.9	25.3	40.8	33.3
3DVG-Transformer [60]	81.9	60.6	39.3	28.4	47.6	34.7
ViL3DRel [7]	81.6	68.6	40.3	30.7	47.9	37.7
3DJCG [4]	83.5	64.3	41.4	30.8	49.6	37.3
D3Net [6]	–	72.0	–	30.1	–	37.9
M3DRef-CLIP [58]	85.3	77.2	43.8	36.8	51.9	44.7
3D-VisTA [65]	81.6	75.1	43.7	39.1	50.6	45.8
3D-LLM (Flamingo) [22]	–	–	–	–	21.2	–
3D-LLM (BLIP2-flant5) [22]	–	–	–	–	30.3	–
Grounded 3D-LLM [9]	–	–	–	–	47.9	44.1
PQ3D [66]	86.7	78.3	51.5	46.2	57.0	51.2
ChatScene [24]	89.6	82.5	47.8	42.9	55.5	50.2
LLaVA-3D [64]	–	–	–	–	54.1	42.2
Video 3D-LLM [62]	88.0	78.3	50.9	45.3	58.1	51.7
3DRS (Ours)	87.4	77.9	57.0	50.8	62.9	56.1

Table 8: Performance comparison on the validation set of Multi3DRefer [58]. ZT: zero-target, ST: single-target, MT: multi-target, D: distractor.

Method	ZT w/o D	ZT w/ D	ST w/o D		ST w/ D		MT		ALL	
	F1	F1	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5
M3DRef-CLIP [58]	81.8	39.4	53.5	47.8	34.6	30.6	43.6	37.9	42.8	38.4
D3Net [6]	81.6	32.5	–	38.6	–	23.3	–	35.0	–	32.2
3DJCG [4]	94.1	66.9	–	26.0	–	16.7	–	26.2	–	26.6
Grounded 3D-LLM [9]	–	–	–	–	–	–	–	–	45.2	40.6
PQ3D [66]	85.4	57.7	–	68.5	–	43.6	–	40.9	–	50.1
ChatScene [24]	90.3	62.6	82.9	75.9	49.1	44.5	45.7	41.1	57.1	52.4
Video 3D-LLM [62]	94.7	78.5	82.6	73.4	52.1	47.2	40.8	35.7	58.0	52.7
3DRS (Ours)	95.6	79.4	79.6	71.4	57.0	51.3	43.0	37.8	60.4	54.9

Table 9: Performance comparison on the validation set of ScanQA [2]. EM indicates exact match accuracy, and B-1, B-2, B-3, B-4 denote BLEU-1, -2, -3, -4, respectively.

Method	EM	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
ScanQA [2]	21.05	30.24	20.40	15.11	10.08	33.33	13.14	64.86
3D-VisTA [65]	22.40	–	–	–	10.40	35.70	13.90	69.60
Oryx-34B [31]	–	38.00	24.60	–	–	37.30	15.00	72.30
LLaVA-Video-7B [59]	–	39.71	26.57	9.33	3.09	44.62	17.72	88.70
3D-LLM (Flamingo) [22]	20.40	30.30	17.80	12.00	7.20	32.30	12.20	59.20
3D-LLM (BLIP2-flant5) [22]	20.50	39.30	25.20	18.40	12.00	35.70	14.50	69.40
Chat-3D [47]	–	29.10	–	–	6.40	28.50	11.90	53.20
NaviLLM [63]	23.00	–	–	–	12.50	38.40	15.40	75.90
LL3DA [8]	–	–	–	–	13.53	37.31	15.88	76.79
Scene-LLM [16]	27.20	43.60	26.80	19.10	12.00	40.00	16.60	80.00
LEO [25]	–	–	–	–	11.50	39.30	16.20	80.00
Grounded 3D-LLM [9]	–	–	–	–	13.40	–	–	72.70
ChatScene [24]	21.62	43.20	29.06	20.57	14.31	41.56	18.00	87.70
LLaVA-3D [64]	27.00	–	–	–	14.50	50.10	20.70	91.70
Video 3D-LLM [62]	30.10	47.05	31.70	22.83	16.17	49.02	19.84	102.06
3DRS (Ours)	30.30	48.37	32.67	23.79	17.22	49.82	20.47	104.78

discrimination is required. These results demonstrate that 3DRS effectively leverages multi-view images for robust spatial understanding and accurate object localization.

Multi3DRefer. In Tab. 8, 3DRS achieves the best overall results on the Multi3DRefer validation set, with top F1 scores in both standard and challenging scenarios. Our method consistently outperforms previous approaches, especially in the difficult zero-target and distractor settings, demonstrating superior robustness and spatial understanding.

ScanQA. In Tab. 9, 3DRS achieves the best performance on the ScanQA validation set across almost all metrics, including EM, BLEU scores, METEOR, and CIDEr, demonstrating its strong effectiveness for 3D question answering.

SQA3D. In Tab. 10, 3DRS achieves the highest scores on the SQA3D test set, outperforming all previous approaches on almost every question type as well as in the overall average, which demonstrates its superior capability for 3D question answering across diverse scenarios.

Scan2cap. In Tab. 11, 3DRS achieves the best performance on the Scan2Cap validation set in terms of CIDEr (C), and remains highly competitive on other metrics such as BLEU-4, METEOR, and ROUGE-L, demonstrating strong overall effectiveness for 3D captioning.

A.4 Ablation Study

Table 12 shows the effect of applying supervision to different numbers of network layers across multiple 3D scene understanding tasks, including object localization, captioning, and question answering. Supervising only the last layer consistently achieves the best performance on all benchmarks. As more intermediate layers are added for supervision, the results degrade. This suggests that multi-layer supervision may over-constrain geometric features and weaken semantic representations, ultimately hindering downstream performance. Future work may explore more advanced strategies to balance geometric and semantic cues.

Table 10: Performance comparison on the test set of SQA3D [32].

Method	Test set						Avg.
	What	Is	How	Can	Which	Others	
SQA3D [32]	31.60	63.80	46.00	69.50	43.90	45.30	46.60
3D-VisTA [4]	34.80	63.30	45.40	69.80	47.20	48.10	48.50
LLaVA-Video[59]	42.70	56.30	47.50	55.30	50.10	47.20	48.50
Scene-LLM [16]	40.90	69.10	45.00	70.80	47.20	52.30	54.20
LEO [25]	–	–	–	–	–	–	50.00
ChatScene [24]	45.40	67.00	52.00	69.50	49.90	55.00	54.60
LLaVA-3D [64]	–	–	–	–	–	–	55.60
Video 3D-LLM [62]	51.10	72.40	55.50	69.80	51.30	56.00	58.60
3DRS (Ours)	54.40	75.20	57.00	72.20	49.90	59.00	60.60

Table 11: Performance comparison on the validation set of Scan2Cap [12].

Method	@0.5			
	C	B-4	M	R
Scan2Cap [12]	39.08	23.32	21.97	44.48
3DJCG [4]	49.48	31.03	24.22	50.80
D3Net [6]	62.64	35.68	25.72	53.90
3D-VisTA [65]	66.90	34.00	27.10	54.30
LL3DA [8]	65.19	36.79	25.97	55.06
LEO [25]	68.40	36.90	27.70	57.80
ChatScene [24]	77.19	36.34	28.01	58.12
LLaVA-3D [64]	79.21	41.12	30.21	63.41
Video 3D-LLM [62]	83.77	42.43	28.87	62.34
3DRS (Ours)	86.11	41.63	28.97	62.29

Table 12: Distillation on multiple layers.

Supervision	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
Last layer	**62.9**	**56.1**	**60.4**	**54.9**	**41.6**	**86.1**	**104.8**	**30.3**	**60.6**
Last layer + last 3rd layer	61.5	54.8	60.1	**54.9**	41.4	84.4	101.4	29.2	60.5
Last layer + last 3rd + last 5th layer	60.5	53.9	59.0	53.8	40.0	81.1	102.9	30.0	59.6

Table 13 reports the impact of different distillation loss functions, including euclidean loss, cosine loss, and their combination, across various 3D scene understanding benchmarks. The results show that all loss types yield very similar performance, indicating that the choice of feature distance metric has limited influence in our setting.

A.5 Qualitative Results

Visualizations Fig. 4 illustrates qualitative results of our method across three tasks: visual grounding, object captioning, and question answering.

For the visual grounding task (top two rows), the model is required to localize objects within a 3D scene based on natural language descriptions. Each example shows the ground truth bounding box (blue), the result from a baseline method (red), and our prediction (green). In both cases, our method's predictions match the ground truth more closely than the baseline, demonstrating improved grounding accuracy.

In the object captioning task (middle two rows), the model generates descriptive captions for specific objects in the scene. The captions from the ground truth, the baseline, and our method are shown alongside their corresponding regions. We also report CIDEr scores to measure caption quality. Our approach produces more accurate and detailed descriptions with significantly higher CIDEr scores compared to the baseline.

For the question answering task (bottom two rows), the model answers questions about the scene. Ground truth answers, baseline outputs, and our results are provided for each question. Red rectangles highlight the visual evidence used by our model to generate the answers. Our method provides correct answers that align with the ground truth, whereas the baseline often fails to do so.

Overall, the visualizations demonstrate that our approach consistently outperforms the baseline across all tasks, delivering more accurate grounding, richer object descriptions, and more reliable answers to visual questions.

Table 13: Performance with different distillation losses.

Supervision	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	C	EM	EM
Euclidean loss	**62.9**	**56.1**	60.4	54.9	41.6	**86.1**	**104.8**	**30.3**	60.6
Cosine loss	62.2	55.5	**60.4**	**55.2**	41.8	85.9	104.5	30.1	**60.7**
Cosine + Euclidean	62.3	55.7	60.3	55.0	**42.1**	85.8	102.7	29.7	60.2

Figs. 5 and 6 provide a visual summary of how our method performs on three challenging 3D scene understanding tasks. These tasks include identifying objects based on language, generating descriptions for specific regions, and answering spatial questions about the scene.

In the visual grounding examples at the top, the model is challenged to find the correct object in a complex 3D environment based on a textual description. The comparison highlights three bounding boxes for each case: blue for the ground truth, red for the baseline, and green for our result. Our predictions consistently align with the intended targets, showing our model's ability to accurately interpret spatial and semantic cues from language.

The object captioning section in the middle presents how each model describes a highlighted object or area. For each instance, the ground truth, baseline output, and our generated caption are shown, along with their respective CIDEr scores. Our model's captions are both more precise and more faithful to the scene's content, as reflected in the higher evaluation scores.

At the bottom, the question answering task demonstrates the model's reasoning abilities within a 3D environment. The figures show the posed question, the correct answer, the baseline's response, and our model's answer. Even for questions that require counting or locating objects, our approach tends to provide accurate answers, often supported by clear visual evidence in the scene.

Altogether, these qualitative results illustrate that our approach delivers more reliable scene understanding across a variety of tasks, outperforming the baseline in both accuracy and descriptive quality.

A.6 Broader Impacts

Positive impacts. The advancement of 3D perception in AI systems holds significant positive societal potential. Enhanced 3D understanding can benefit applications such as assistive robotics for the elderly and disabled, safer autonomous navigation, improved medical imaging, and immersive educational tools. These technologies have the capacity to improve quality of life, boost accessibility, and enable new forms of human-computer interaction.

Negative impacts. However, the adoption of enhanced 3D perception also raises important privacy concerns, especially in surveillance and monitoring contexts where individuals' activities or environments could be reconstructed and analyzed without their consent. To address these risks, it is crucial to apply robust data anonymization methods—such as blurring faces or removing identifiable features—ensure informed consent from data subjects, enforce strict access controls and data security protocols, and adhere to relevant privacy regulations to protect individual rights.

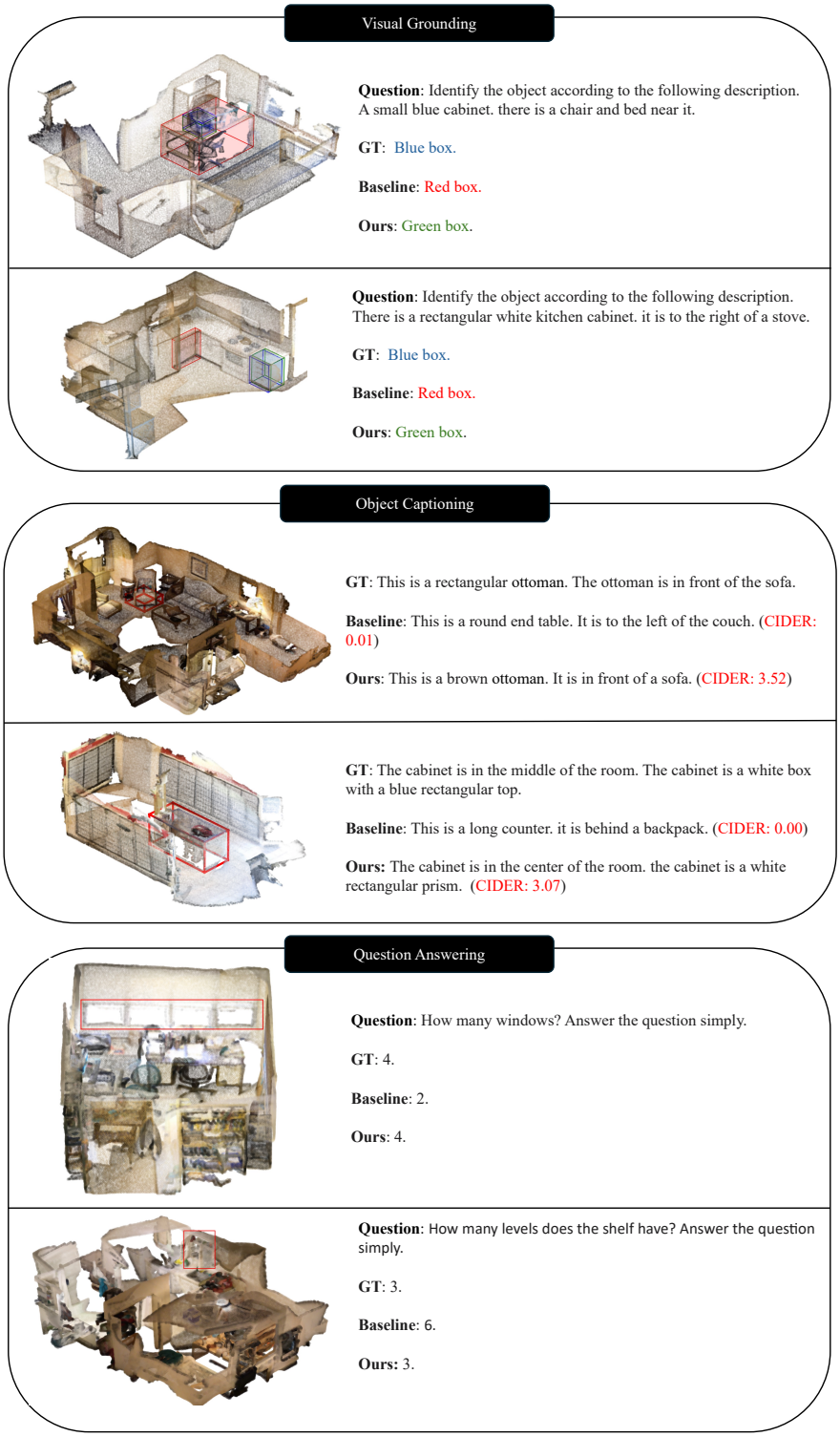


Figure 4: **Visualization of Results Across Different Tasks.** (a) Visual Grounding: The predicted bounding box closely aligns with the ground truth. (b) Object Captioning: Our method generates accurate captions for each referred object. (c) Question Answering: The model provides precise answers, where we use the red rectangles to indicate the visual cues utilized for each response. Best viewed when zoomed in.

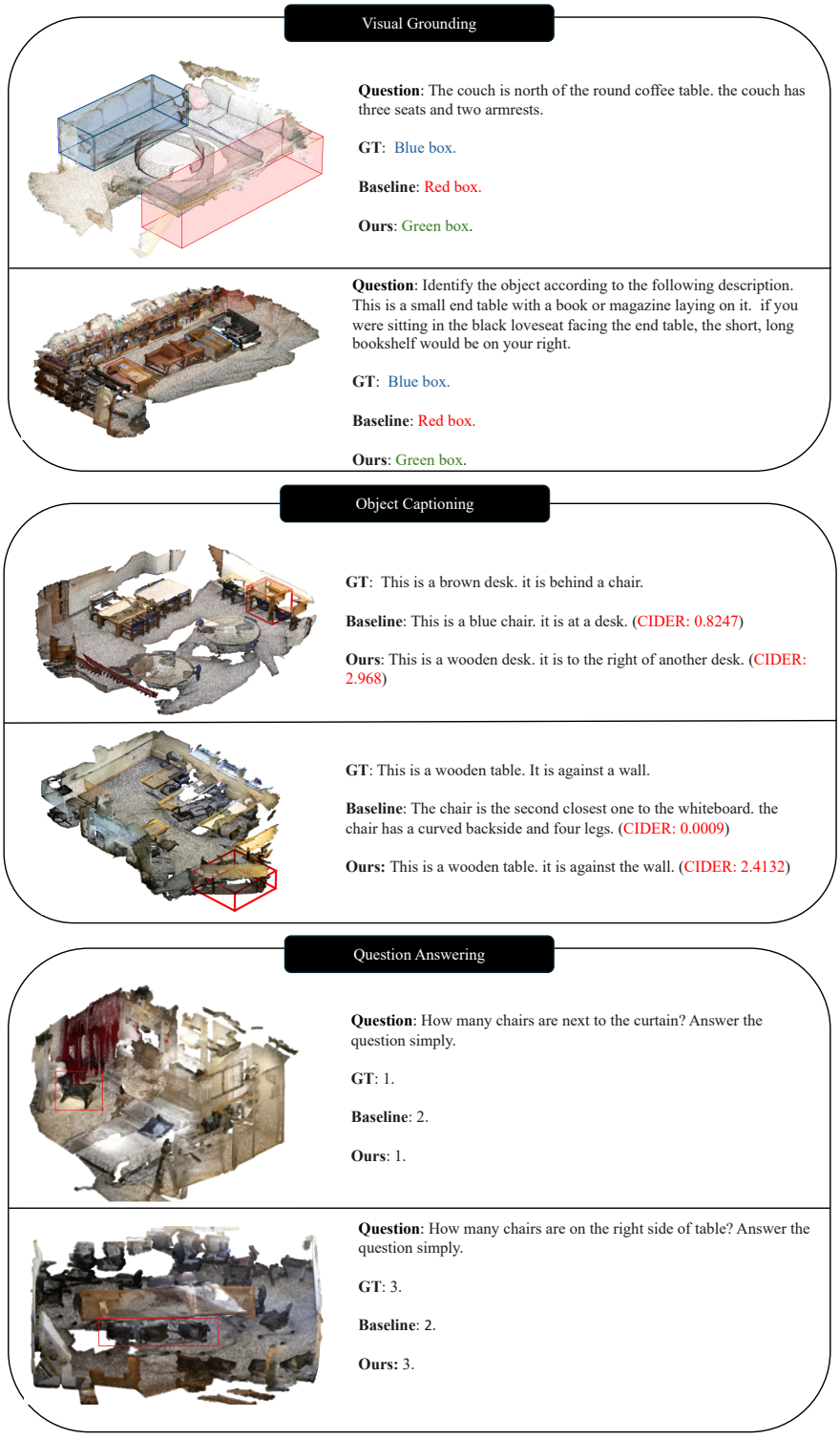


Figure 5: **Visualization of Results Across Different Tasks.** (a) Visual Grounding: The predicted bounding box closely aligns with the ground truth. (b) Object Captioning: Our method generates accurate captions for each referred object. (c) Question Answering: The model provides precise answers, where we use the red rectangles to indicate the visual cues utilized for each response. Best viewed when zoomed in.

Figure 6: **Visualization of Results Across Different Tasks.** (a) Visual Grounding: The predicted bounding box closely aligns with the ground truth. (b) Object Captioning: Our method generates accurate captions for each referred object. (c) Question Answering: The model provides precise answers, where we use the red rectangles to indicate the visual cues utilized for each response. Best viewed when zoomed in.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (*e.g.*, "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope about 3D-aware representation learning in MLLMs for scene understanding.

   Guidelines:

   - The answer NA 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 NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

Justification: The limitation section is provided in Sec. 6.

Guidelines:

- The answer NA 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: [NA]

   Justification: This paper does not contain theoretical assumptions and proofs.

   Guidelines:

   - The answer NA 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: The detailed method design is given in Sec. 2, and experimental details are all provided in Sec. 3 and App. A.2.

   Guidelines:

- The answer NA 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 the supplemental material?

   Answer: [No]

   Justification: We will publicly release the code and related instructions in the near future.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the necessary details are provided in Sec. 3 and App. A.2.

   Guidelines:

   - The answer NA 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: Our experimental settings are consistent with those used in prior works published at top conferences, which do not report error bars as well. Besides, due to resource constraints, we were unable to perform the multiple runs required to report error bars.

   Guidelines:

   - The answer NA 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: The information of GPUs we use for experiments is provided in Sec. 2.

   Guidelines:

   - The answer NA 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 strictly follow the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss the broader impacts in App. A.6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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.*, pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper uses all public datasets for training models and has no such risk.

Guidelines:

- The answer NA 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: The original papers of the datasets are cited, and their licenses are provided in the reference.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We detail the training and inference settings in Sec. 3 and App. A.2. As for the code, we will release it in the near future.

    Guidelines:

    - The answer NA 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: [NA]

Justification: We do not involve human subjects in our experiments.

Guidelines:

- The answer NA 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: [NA]

Justification: We do not involve human subjects in our experiments.

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only use LLM to improve the wording of the paper and implement code for visualizations.

Guidelines:

- The answer NA 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 (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.