Beyond Single Expert: Harmonizing Diverse Visual Priors in MLLMs for Spatial Understanding

Xiao Lin **Xiaohu Huang** **Kai Han***
The University of Hong Kong
{lllinxiao, huangxiaohu}@connect.hku.hk, kaihanx@hku.hk

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated substantial promise in spatial understanding. Existing works typically incorporate prior knowledge extracted from a pre-trained foundation model to further enhance the spatial awareness of MLLMs. In this paper, we first reveal that when integrating diverse foundation models into MLLMs, different models provide complementary spatial priors that benefit different tasks. Motivated by this, we propose **ViPS**, a novel multi-model prior framework designed to fully unleash the potential of incorporating multiple **Vi**sual **P**riors from diverse models into MLLMs for **S**patial understanding. Specifically, ViPS introduces an Efficient Prior Proxy to generate multiple foundational priors with minimal inference overhead, and a Dynamic Prior Fusion mechanism to achieve harmonious and context-aware prior fusion and injection from the prior proxies. Extensive experiments demonstrate that ViPS successfully harmonizes diverse visual priors, establishing new state-of-the-art performance across multiple complex spatial reasoning and 3D spatial understanding benchmarks. Project page: https://visual-ai.github.io/vips

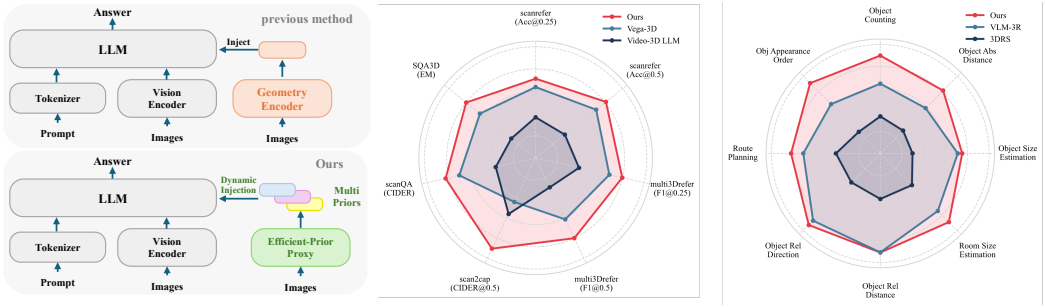


Figure 1: **Comparison of Existing Single-Expert Paradigms and Our Multi-Prior Framework. Top-left:** Existing paradigms typically rely on a single external encoder (e.g., VGGT) to provide visual priors for MLLMs. **Bottom-left:** In contrast, our approach integrates diverse knowledge from multiple expert models into the MLLM. **Middle and right:** Extensive evaluations demonstrate that our method achieves state-of-the-art performance across multiple benchmarks.

1 Introduction

Spatial understanding serves as the foundation for real-world reasoning and interaction, acting as a pivotal technology for critical applications such as robotic navigation and embodied intelligence.

*Corresponding author.

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

The rapid advancement of Multimodal Large Language Models (MLLMs)[Wang et al., 2024, Li et al., 2024, Chen et al., 2024d, Hurst et al., 2024, Team et al., 2024, Bai et al., 2025] has brought substantial breakthroughs to this domain. By bridging the general reasoning capabilities of Large Language Models (LLMs) with visual or point-cloud encoders [Xu et al., 2024, Qi et al., 2024], MLLMs facilitate sophisticated spatial perception and reasoning within complex environments.

Recently, numerous studies have explored incorporating knowledge from pre-trained foundation models (e.g., VGGT [Wang et al., 2025b], WAN [Wan et al., 2025]) into MLLMs [Huang et al., 2025, Wu et al., 2026, Li et al., 2026, Zheng et al., 2025a]. Typically, this paradigm involves leveraging foundational features as auxiliary inputs or aligning the latent representations of MLLMs with foundation models through distillation or feature injection. Such a paradigm has proven effective in augmenting MLLMs with domain-specific expert knowledge to boost spatial awareness. Despite their success, existing efforts are predominantly restricted to exploiting prior knowledge from a single expert, while overlooking the potential of fusing diverse knowledge from multiple foundation models. Intuitively, different foundation models capture distinct priors shaped by their varied pre-training data and objectives, making it highly desirable to aggregate their respective advantages in MLLMs. Therefore, how to synergistically harmonize these complementary strengths from distinct models remains a significant, yet unresolved challenge.

In this paper, we seek to explore two questions that form our core motivations: (i) *Do different foundation models exhibit distinct specializations across various spatial understanding tasks when serving as priors for MLLMs?* (ii) *How can we synergistically integrate the expert knowledge from multiple foundation models into MLLMs?* Regarding the first question, we conduct comprehensive empirical studies by integrating the priors of different foundation models [Wang et al., 2025b, Lin et al., 2025, Liu et al., 2025a, Wan et al., 2025, Heinrich et al., 2025] into MLLMs across a wide range of spatial understanding tasks [Yang et al., 2025, Chen et al., 2020, Zhang et al., 2023b, Chen et al., 2021, Azuma et al., 2022, Ma et al., 2022]. Our results reveal that these models contribute differently to specific tasks, indicating a strong complementarity among diverse visual priors and underscoring the necessity of multi-model integration. Based on these findings, a straightforward solution to the second question would be a naive ensemble of multiple priors. However, such an approach suffers from several critical drawbacks: First, the computational overhead becomes prohibitive, as extracting priors from each foundation model requires an independent forward pass during inference. Second, since different visual priors inherently exhibit distinct strengths and specializations, a naive ensemble not only fails to prioritize the most relevant prior knowledge for a given task, but also risks confusing the MLLM due to the distribution disparities among different foundation models.

To address these challenges, we propose **ViPS** (illustrated in Figure 1), a novel multi-model prior framework designed to fully unleash the potential of multiple **Vi**sual **P**riors in MLLMs for **S**patial understanding. ViPS harmoniously integrates priors from multiple models into the MLLM by introducing an Efficient Prior Proxy and a Dynamic Prior Fusion mechanism. Specifically, the Efficient Prior Proxy is designed to generate multiple foundational priors without requiring multiple forward passes of foundation models. Instead of directly deploying all foundation models independently, we employ one base model alongside several lightweight proxies to estimate the priors of the foundation models and align the proxy outputs with the ground-truth priors during training. This design is motivated by the insight that different foundation models often share common low- and mid-level visual and geometric features. Therefore, a robust base model providing sufficient representations is enough, from which other distinct priors can be effectively distilled using lightweight proxies. Furthermore, the Dynamic Prior Fusion is proposed to achieve harmonious and context-aware prior selection from proxies. To this end, we first generate dynamic weights based on the input task, and then use these weights to aggregate the priors from different foundation models into a fused multi-expert prior, which is subsequently injected into the MLLM. Additionally, a set of zero-initialized convolutional layers is applied before aggregation to ensure that diverse priors are harmoniously fused. Extensive experiments demonstrate that ViPS achieves state-of-the-art performance across diverse spatial reasoning and spatial understanding benchmarks.

To summarize, we make the following contributions: *First*, we conduct an empirical study on integrating diverse foundation models into MLLMs, revealing that different visual priors exhibit distinct specializations across various tasks. *Second*, we propose **ViPS**, a novel multi-model prior framework for spatial understanding, which features an Efficient Prior Proxy to generate multi-model priors with minimal overhead, and a Dynamic Prior Fusion mechanism for harmonious and

context-aware prior integration. *Third*, extensive experiments demonstrate that **ViPS** achieves new state-of-the-art performance across multiple spatial reasoning and spatial understanding benchmarks.

2 Method

2.1 Empirical Study on Diverse Model Priors

Preliminary. Our work focuses on spatial understanding based on MLLMs. Formally, given a visual input comprising a sequence of frames (e.g., a video clip or multi-view images of a scene) $\mathcal{V} = \{v_1, v_2, \ldots, v_T\} \in \mathbb{R}^{T\times H\times W\times 3}$, alongside a corresponding textual task instruction $\mathcal{X}_{text} = \{x_1, x_2, \ldots, x_N\}$, the objective of the MLLM is to generate a coherent and accurate textual response $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$ that adequately answers the query by reasoning over the holistic spatial structures of the scene. In standard video-based MLLM architectures (e.g., LLaVA [Li et al., 2024], Qwen2-VL [Wang et al., 2024]), the visual sequence $\mathcal{V}$ is first processed by a generic vision encoder Φ_{vis}, producing visual embeddings $F_{vis} = \Phi_{vis}(\mathcal{V})$. A modality projector $\mathcal{P}$ then aligns these embeddings into the text semantic space. The LLM, denoted as f_{LLM}, takes the concatenated tokens of the projected visual features and the text instruction to auto-regressively predict the response:

$$P(\mathcal{Y} \mid \mathcal{V}, \mathcal{X}_{text}) = \prod_{i=1}^{M} P\left(y_i \mid \mathcal{P}(F_{vis}), \mathcal{X}_{text}, y_{<i}; \theta\right), \tag{1}$$

where θ encapsulates the learnable parameters of the system. To bridge the gap in 3D spatial awareness, recent approaches explicitly introduce external foundation model priors. Given a pre-trained visual expert $\mathcal{E}$ (e.g., encoders from foundation models like VGGT), it extracts domain-specific prior features $F_{prior} = \mathcal{E}(\mathcal{V})$. The LLM leverages this supplementary knowledge to enhance generation, yielding $P(\mathcal{Y} \mid F_{vis}, F_{prior}, \mathcal{X}_{text})$. While integrating a single prior model provides distinct utility, relying exclusively on an individual $\mathcal{E}$ inevitably limits the representation capacity, forming the motivation for our extensive empirical investigations.

Key Finding and Motivation. As highlighted in the introduction, existing efforts are restricted to exploiting the priors of a single model, overlooking the potential of fusing diverse knowledge from multiple foundation models. Fundamentally, different foundation models encapsulate distinct representations and spatial semantics, thereby endowing them with disparate advantages when serving as priors for MLLMs. To demonstrate this, we select a diverse set of foundation models for evaluation, including VGGT [Wang et al., 2025b], DepthAnything3 [Lin et al., 2025], TraceAnything [Liu et al., 2025a], Wan2.1 [Wan et al., 2025], and RADIO [Heinrich et al., 2025]. Their feature distributions and activation patterns are analyzed in Appendix C. Subsequently, we explore their relative advantages across a wide range of spatial understanding tasks when serving as priors for MLLMs (See A.3 for more details). The results are shown in Figure 2. From a column-wise perspective, the best-performing model varies across different sub-tasks. From a row-wise perspective, each specific model

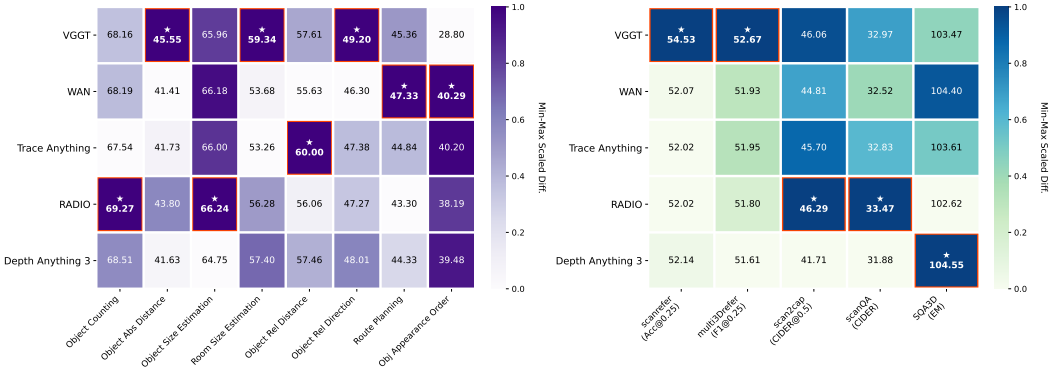


Figure 2: **Relative Performance of Diverse Foundation Models.** We evaluate the performance of various foundation models when serving as priors for MLLMs across a wide range of spatial understanding tasks. Notably, no single model dominates all metrics, thoroughly motivating the need for the multi-model prior integration.

achieves its optimal performance on disparate sub-tasks. These results demonstrate that no single foundation model achieves universal dominance; instead, different models exhibit distinct strengths and specializations. The highly complementary nature of these diverse visual priors directly motivates our core objective: to efficiently extract and adaptively harmonize distinct prior knowledge from multiple foundation models for comprehensive spatial reasoning.

2.2 Proposed Framework: ViPS

Building upon the above insights, we introduce **ViPS** (**Vi**sual **P**riors for **S**patial understanding), a novel framework designed to fully unleash the potential of leveraging diverse prior models in MLLMs. The overall framework is illustrated in Figure 3. ViPS first processes the input visual sequence $\mathcal{V}$ through a standard vision encoder Φ_{vis} [Zhai et al., 2023], producing the base visual embeddings F_{vis}. Instead of relying on a single expert or naively ensembling multiple heavy models, ViPS seamlessly integrates comprehensive prior knowledge through two pivotal mechanisms: **Efficient Prior Proxy** and **Dynamic Prior Fusion**.

Specifically, the Efficient Prior Proxy utilizes a base foundation model alongside lightweight MLPs to efficiently estimate the prior features of various external models, generating a set of diverse prior representations $\{F^1_{prior}, F^2_{prior}, \ldots, F^K_{prior}\}$. Subsequently, these representations are fed into the Dynamic Prior Fusion part, which first applies independent, zero-initialized convolutional layers [Zhang et al., 2023a] to each prior branch and then employs a context-aware weighting mechanism, guided by the final token of the input instruction $\mathcal{X}_{text}$, to dynamically compute fusion weights. Finally, the harmoniously fused multi-expert prior $\hat{F}_{prior}$ is incorporated into the MLLM, empowering the LLM to leverage the most relevant visual knowledge tailored to the specific spatial reasoning task. We detail the formulations of these two components below.

Efficient Prior Proxy. Our framework aims to effectively integrate highly complementary priors from various foundation models for comprehensive spatial understanding. Directly extracting visual features from a set of K distinct foundation models, denoted as $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_K\}$, conventionally requires K independent forward passes, i.e., computing $\mathcal{E}_k(\mathcal{V})$ for $k \in \{1, \ldots, K\}$. This paradigm introduces a severe computational bottleneck, as the inference latency and memory footprint scale linearly with the number of integrated priors. To decouple the computational overhead from the integration of multiple models while effectively preserving the rich diversity of visual prior knowledge, we propose the efficient prior proxy mechanism. Specifically, we use a single robust vision encoder as the base model $\mathcal{E}_{base}$ to extract a unified foundational representation:

$$F_{base} = \mathcal{E}_{base}(\mathcal{V}) \in \mathbb{R}^{S \times D_{base}}, \tag{2}$$

where S and D_{base} denote the sequence length and channel dimension of the visual tokens, respectively. In our experiments, the base model is directly initialized with the encoder of one of the foundation models (we present experiments using encoders from different foundation models as the base model in the appendix).

To approximate the distinct knowledge of the K targeted foundation models, we instantiate a set of K lightweight proxy networks. Each proxy $\phi_k(\cdot)$ is implemented as a simple Multi-Layer Perceptron (MLP) designed to extract specific prior semantics (e.g., depth cues, geometric boundaries, or detailed semantics) from the shared foundational feature F_{base}. The representation for the k-th prior is uniformly formulated as:

$$F^k_{prior} = \phi_k(F_{base}) \in \mathbb{R}^{S \times D_{prior}}, \tag{3}$$

where D^k_{prior} is the feature dimension for the k-th prior feature.

To guarantee the fidelity of these estimated prior representations, we explicitly supervise the output of prior proxies during the training phase. Specifically, the features extracted by the K distinct foundation models are utilized as the ground-truth targets, denoted as $F^k_{gt} = \mathcal{E}_k(\mathcal{V})$. We apply an L_2 loss to enforce strict alignment between the proxy outputs F^k_{prior} and their corresponding ground-truth priors F^k_{gt}:

$$\mathcal{L}_{alignment} = \sum_{k=1}^{K} \left\| F^k_{prior} - F^k_{gt} \right\|_2^2. \tag{4}$$

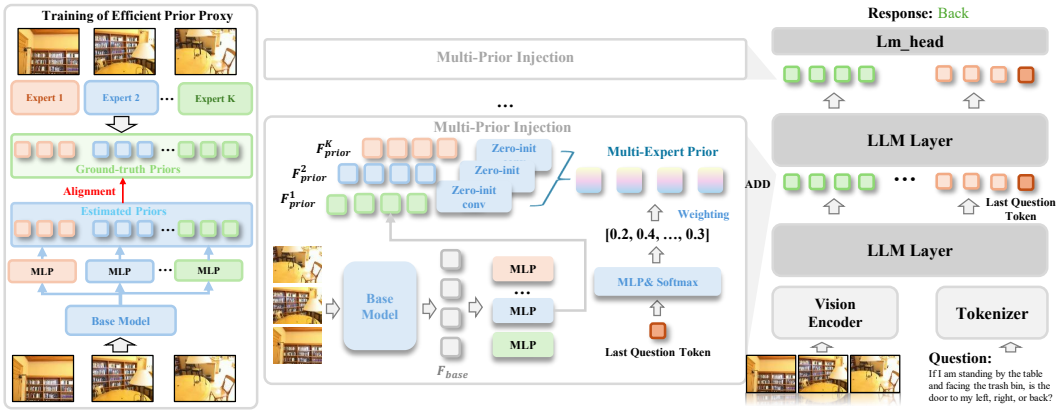


Figure 3: **Overview of the Proposed ViPS Framework.** The framework integrates distinct prior knowledge from multiple foundation models via the Efficient Prior Proxy and coordinates them using Dynamic Prior Fusion for comprehensive spatial reasoning.

The viability of this proxy-based estimation stems from the intuition that different foundation models inherently share substantial low- and mid-level visual semantics; thus, a robust base representation contains sufficient foundational knowledge from which distinct, high-level prior semantics can be efficiently derived via shallow proxies. Given that the computational cost of a shallow MLP ϕ_k is negligible compared to a full foundation model, our framework can scale to an arbitrary number of diverse priors without linearly increasing inference overhead, while ensuring that the estimated visual priors remain accurate through the alignment loss.

Dynamic Prior Fusion. With the diverse prior representations $\{F^1_{prior}, \ldots, F^K_{prior}\}$ efficiently established, the remaining challenge is seamlessly integrating them into the MLLM. Directly adding these distinct and heterogeneous priors fails to prioritize the most effective priors for the specific input task, and can overwhelm the language model, consequently degrading overall performance. To achieve harmonious and context-aware integration, we propose the Dynamic Prior Fusion mechanism.

Accordingly, we design a dynamic weighting mechanism to selectively focus on the precise knowledge required by the given task. Specifically, we extract the representation of the final token of the text query, denoted as x^{text}_{last}, which naturally encapsulates the aggregated semantics of the entire instruction due to the causal mechanism of the LLM. This context vector is passed through an MLP to compute a set of K-dimensional logits $z \in \mathbb{R}^K$:

$$z = \text{MLP}_{weight}(x^{text}_{last}). \tag{5}$$

A Softmax activation is subsequently applied to z to obtain the normalized fusion weights $w \in [0, 1]^K$. These weights will be utilized to dynamically combine the diverse priors. However, our experiments show that directly doing so fails to yield progressive performance improvements. We attribute this phenomenon to the disparate prior distributions originating from different source models, which tend to confuse the MLLM during the early stages (see Appendix C and Section 3.4). To overcome this issue and enable progressive prior injection, we apply an independent zero-initialized convolutional layer, denoted as $\text{ZeroConv}_k(\cdot)$, to each proxy branch prior to the weighted fusion. This strategy ensures that the outputs of the convolution are initially zero, preserving the original MLLM representations and preventing the prior perturbations from disrupting early training, while gradually scaling up as the MLLM progressively learns from the diverse prior knowledge. The processed prior $\tilde{F}^k_{prior}$ is computed as:

$$\tilde{F}^k_{prior} = \text{ZeroConv}_k(F^k_{prior}). \tag{6}$$

Finally, the context-aware weights are utilized to linearly combine the aligned priors, yielding the overall harmonized multi-expert prior $\hat{F}_{prior}$:

$$\hat{F}_{prior} = \sum_{k=1}^{K} w_k \cdot \tilde{F}^k_{prior}. \tag{7}$$

This fused representation $\hat{F}_{prior}$, which dynamically captures the most pertinent visual prior requested by the query, is directly added element-wise to the corresponding projected visual embeddings $\mathcal{P}(F_{vis})$ (i.e., the image tokens) of the MLLM. In this way, Dynamic Prior Fusion seamlessly integrates diverse priors into the MLLM, adaptively tailoring knowledge selection to the input task while leveraging zero-initialization to ensure progressive learning without disrupting early training.

It is worth noting that, for simplicity of presentation, the above formulation describes the prior injection process at a single layer. In practice, we uniformly apply this Dynamic Prior Fusion mechanism across five layers of the MLLM. This multi-layer injection ensures that the diverse prior knowledge is deeply integrated, progressively guiding the spatial reasoning of the model.

3 Experiments

3.1 Dataset and Evaluation Metric

We evaluate our method across a total of six datasets: VSI-Bench [Yang et al., 2025] for spatial reasoning tasks, and five ScanNet-series benchmarks for 3D spatial understanding (ScanRefer [Chen et al., 2020], Multi3DRefer [Zhang et al., 2023b], Scan2Cap [Chen et al., 2021], ScanQA [Azuma et al., 2022], and SQA3D [Ma et al., 2022]). Specifically, VSI-Bench comprises eight fine-grained spatial reasoning tasks: object counting, absolute distance estimation, object size estimation, room size estimation, relative distance, relative direction, route planning, and object appearance order. Following the standard protocol, we report the accuracy (%) across all its sub-tasks. The five ScanNet-series benchmarks, derived from the ScanNet corpus, span three primary tasks: (i) *3D visual grounding*, where ScanRefer targets free-form object localization and Multi3DRefer extends this to multi-target and zero-target ambiguity; (ii) *dense captioning*, using Scan2Cap to generate natural language descriptions for localized objects; and (iii) *embodied question answering*, employing ScanQA for geometry-grounded open-ended questions and SQA3D for complex situated reasoning from a specific agent perspective. Following standard protocols, we report Acc@0.25/0.5 for ScanRefer, F1@0.25/0.5 for Multi3DRefer, CIDEr@0.5 for Scan2Cap, CIDEr and Exact Match (EM) for ScanQA, and Exact Match for SQA3D. More details can be found in the Appendix.

3.2 Implementation Detail

We use both Qwen2-VL-7B [Wang et al., 2024] and Qwen3-VL-8B [Bai et al., 2025] as base MLLMs on VSI-Bench; for all other experiments, we adopt Qwen2-VL-7B as the default model. To comprehensively capture spatial knowledge, we integrate $K = 5$ distinct foundation models to extract complementary visual priors. Specifically, we employ VGGT [Wang et al., 2025b] for general visual-geometric feature extraction, which also serves as the base foundation model in the Efficient Prior Proxy. We further incorporate DepthAnything3 [Lin et al., 2025] for precise monocular depth cues, TraceAnything [Liu et al., 2025a] for object-centric motion and relationship tracking, Wan2.1 [Wan et al., 2025] for rich spatio-temporal dynamics, and RADIO [Heinrich et al., 2025] for robust generic visual representations. These diverse priors collectively provide a holistic understanding of 3D scenes. We uniformly sample 32 frames from each video as the input to the vision encoder. During training, we first apply the alignment loss to train the Efficient Prior Proxy, and then freeze the proxies while fine-tuning the MLLM. We freeze the vision encoder and apply Low-Rank Adaptation (LoRA) [Hu et al., 2022] to the LLM backbone during MLLM fine-tuning. We optimize the model using the Adam optimizer [Kingma and Ba, 2014] with a batch size of 16 and a maximum learning rate of 1×10^{-5}. The model is trained for 1 epoch on each dataset. More details can be found in the Appendix.

3.3 Comparison with State-of-the-Art Methods

We evaluate ViPS against leading MLLMs and specialized 3D spatial understanding models. Table 1 and Table 2 summarize our results on VSI-Bench and the ScanNet-series benchmarks respectively.

Spatial Reasoning on VSI-Bench. As shown in Table 1, ViPS achieves a leading average score of 63.8%, surpassing the previous best spatial-enhanced model VLM-3R [Fan et al., 2025] (57.2%). While trailing VG-LLM-8B [Zheng et al., 2025a] in Appearance Order, ViPS dominates most other categories, proving the advantage of dynamically injecting multiple priors over relying solely on a single foundation model like VLM-3R or 3DRS [Huang et al., 2025].

Table 1: **Performance Comparison on VSI-Bench.** ViPS achieves a leading average score across the eight spatial reasoning sub-tasks, surpassing prior spatial-enhanced MLLMs.

		Numerical Answer				Multiple-Choice Answer			
Model	Avg.	Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Proprietary Models (API)									
GPT-4o [Hurst et al., 2024]	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5-Pro [Team et al., 2024]	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-source Models									
LongVA-7B [Zhang et al., 2024]	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7
LongVILA-8B [Chen et al., 2024c]	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5
InternVL2-8B [Chen et al., 2024d]	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B [Chen et al., 2024d]	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
VILA-1.5-40B [Liu et al., 2025b]	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
LLaVA-OneVision-7B [Li et al., 2024]	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B [Li et al., 2024]	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-NeXT-Video-7B [Liu et al., 2024]	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B [Liu et al., 2024]	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Spatial-Enhanced Models									
vsGRPO-V-7B [Liao et al., 2025]	40.7	59.9	29.6	50.8	48.3	35.4	35.6	34.0	31.5
SPAR-8B [Zhang et al., 2025]	41.1	-	-	-	-	-	-	-	-
SpaceR-7B [Ouyang et al., 2025]	45.6	-	-	-	-	-	-	-	-
VG-LLM-4B [Zheng et al., 2025a]	45.9	65.6	37.4	54.8	60.2	42.3	46.3	33.0	25.9
VG-LLM-8B [Zheng et al., 2025a]	50.1	67.2	38.0	59.3	63.2	47.0	43.9	33.0	49.4
3DRS-7B [Huang et al., 2025]	45.9	68.7	34.8	53.6	56.6	40.9	43.2	30.4	39.2
Vega-3D [Wu et al., 2026]	50.5	69.7	35.9	58.0	60.8	45.1	43.1	30.9	**60.5**
VLM-3R [Fan et al., 2025]	57.2	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
ViPS (qwen2-vl-7b)	**63.8**	71.5	**60.8**	70.7	**71.6**	**65.5**	**84.7**	**51.0**	34.3
ViPS (qwen3-vl-8b)	**63.8**	**82.6**	50.5	**81.4**	64.1	62.8	64.1	46.2	**58.3**

3D spatial understanding on ScanNet-series. Table 2 shows our results across grounding, captioning, and QA tasks. Compared to generalist 3D-LLMs like Vega-3D [Wu et al., 2026] and 3DRS [Huang et al., 2025], ViPS delivers highly competitive performance. It yields top scores on visual grounding (ScanRefer, 64.6% Acc@0.25) and QA (ScanQA, 107.9 CIDEr), outperforming Vega-3D. Although ViPS slightly trails 3DRS on Scan2Cap, these results confirm that our unified proxy injection effectively harnesses multiple foundation priors for complex 3D spatial understanding tasks.

3.4 Ablation Study

Effectiveness of Individual Visual Priors. Table 3 evaluates the impact of individual foundation models compared to a baseline MLLM without prior injection. While each single prior yields distinct performance gains, our full ViPS framework integrating all five priors achieves the highest scores across all metrics. This confirms both the effectiveness of individual priors and the strong complementarity of harmonizing heterogeneous visual knowledge for 3D spatial understanding.

Effectiveness of Dynamic Prior Injection. Table 4 compares the full ViPS model against two injection variants. First, replacing zero-initialized convolutions with standard random initialization (*w/o Zero-init*) causes a severe performance drop. This confirms our motivation that zero-init is essential to prevent the diverse, unaligned prior distributions from confusing the MLLM during early training, preserving the original latent space and enabling progressive prior injection. Second, substituting the dynamic proxy fusion with a straightforward feature sum (*Vanilla Addition*) degrades performance. This demonstrates that our dynamic mechanism, which adapts to the input query, achieves more effective prior injection than static feature aggregation.

Effectiveness of Efficient Prior Proxy. To demonstrate the efficiency and accuracy of our Efficient Prior Proxy, we compare our method against an upper-bound setting (*w/ GT Priors*) where ground-truth features from all five foundation models are used to replace priors from the Efficient Prior Proxy. As shown in Table 5, our method employs multiple lightweight proxies to replace full foundation models, significantly reducing parameter overhead and inference latency ($1\times$ vs. $\sim 5\times$) with only a

Table 2: **Performance Comparison on ScanNet-Series Benchmarks.** ViPS delivers competitive performance across 3D visual grounding (ScanRefer, Multi3DRefer), dense captioning (Scan2Cap), and embodied question answering (ScanQA, SQA3D).

Method	ScanRefer		Multi3DRefer		Scan2Cap	ScanQA		SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	C@0.5	CIDEr	EM	EM
Specialists								
ScanRefer [Chen et al., 2020]	37.3	24.3	–	–	–	–	–	–
M3DRef-CLIP [Zhang et al., 2023b]	51.9	44.7	42.8	–	38.4	–	–	–
Scan2Cap [Chen et al., 2021]	–	–	–	–	35.2	–	–	–
ScanQA [Azuma et al., 2022]	–	–	–	–	–	64.9	21.1	47.2
3D-VisTA [Zhu et al., 2023]	50.6	45.8	–	–	66.9	69.6	22.4	48.5
Generalists								
3D-LLM (BLIP2-flant5) [Hong et al., 2023]	30.3	–	–	–	–	69.4	20.5	–
Chat-3D v2 [Huang et al., 2023]	42.5	38.4	45.1	41.6	63.9	87.6	–	54.7
SceneLLM [Fu et al., 2024]	–	–	–	–	–	80.0	27.2	53.6
Grounded 3D-LLM [Chen et al., 2024b]	47.9	44.1	45.2	40.6	70.6	72.7	–	–
PQ3D [Zhu et al., 2024b]	57.0	51.2	–	50.1	80.3	–	–	47.1
ChatScene [Huang et al., 2023]	55.5	50.2	57.1	52.4	77.1	87.7	21.6	54.6
LLaVA-3D [Zhu et al., 2024a]	54.1	42.4	–	–	79.2	91.7	27.0	55.6
Inst3D-LLM [Yu et al., 2025]	57.8	51.6	58.3	53.5	79.7	88.6	24.6	–
3D-LLaVA [Deng et al., 2025]	51.2	40.6	–	–	78.8	92.6	–	54.5
Video-3D LLM [Zheng et al., 2025b]	58.1	51.7	58.0	52.7	83.8	102.1	30.1	58.6
3DRS [Huang et al., 2025]	62.9	56.1	60.4	54.9	**86.1**	104.8	30.3	60.6
Vega-3D [Wu et al., 2026]	63.2	56.2	60.8	55.1	83.2	106.3	30.4	61.3
ViPS	**64.6**	**57.6**	**62.0**	**56.5**	85.5	**107.9**	**31.6**	**62.5**

Table 3: **Ablation on Individual Visual Priors.** Each foundation-model prior yields distinct gains over the no-prior baseline, and combining all five priors achieves the best results across the ScanNet-series benchmarks.

Method	ScanRefer		Multi3DRefer		Scan2Cap	ScanQA		SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	C@0.5	C	EM	EM
Baseline	62.1	54.6	59.6	54.4	81.4	104.6	30.5	60.8
+ RADIO	62.9	56.0	61.2	55.5	82.6	105.1	30.7	61.1
+ DepthAnything3	62.9	56.2	61.0	55.7	81.8	106.2	30.9	61.2
+ TraceAnything	62.8	55.8	60.6	55.2	82.7	106.7	30.8	61.4
+ VGGT	63.2	56.3	61.1	55.8	81.9	106.6	31.1	61.6
+ Wan2.1	62.7	55.8	61.1	55.5	81.3	105.5	30.6	61.2
ViPS (Full)	**64.6**	**57.6**	**62.0**	**56.5**	**85.5**	**107.9**	**31.6**	**62.5**

marginal drop in overall performance. We also report the estimation error (0.252 in cosine distance) between the estimated priors and the ground-truth priors during inference. These experimental results consistently demonstrate the effectiveness of our proposed approach. Furthermore, we conduct an ablation study by removing the alignment loss ($\mathcal{L}_{alignment}$) while retaining all other prior proxy structures (in this setting, the whole framework, including the MLLM and Efficient Prior Proxy, is trained end-to-end). Surprisingly, although the performance is inferior to the setting with the alignment loss, it still achieves considerable improvements compared to the baseline. We attribute this to the fact that even without the alignment loss, different prior proxies can still serve as projectors onto distinct sub-feature spaces, enabling the MLLM to learn diverse knowledge from the base model, akin to the multi-head mechanism in attention networks.

3.5 Analysis of Dynamic Prior Weights

To further understand the behavior of our Dynamic Prior Fusion mechanism, we visualize the learned fusion weights (w_k in Equation 7) across different spatial understanding tasks. As illustrated in Figure 4, the weight distribution adapts dynamically depending on different downstream tasks (e.g., ScanQA, ScanRefer, Multi3DRefer, SQA3D, and Scan2Cap), as well as across distinct question types

Table 4: **Ablation on the Dynamic Prior Injection.** Both the zero-initialized convolution and the context-aware dynamic fusion are essential for harmonizing diverse priors without disrupting early training.

Method	ScanRefer		Multi3DRefer		Scan2Cap	ScanQA		SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	C@0.5	C	EM	EM
Baseline	62.1	54.6	59.6	54.4	81.4	104.6	30.5	60.8
w/o Zero-init	63.1	56.1	61.0	55.6	76.4	104.7	30.8	61.1
Vanilla Addition	64.2	57.4	61.2	56.0	82.6	106.4	30.9	61.9
ViPS (Full)	**64.6**	**57.6**	**62.0**	**56.5**	**85.5**	**107.9**	**31.6**	**62.5**

Table 5: **Ablation on the Efficient Prior Proxy.** The lightweight proxies match the upper bound (w/ GT Priors , where features are extracted via independent forward passes of the original foundation models) with only a marginal performance drop, while reducing parameter and inference cost from $\sim 5\times$ to $1\times$.

Method	Efficiency			ScanRefer		Multi3DRefer		S2C	ScanQA		SQA
	Param.	Latency	Err. ↓	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	C@0.5	C	EM	EM
w/ GT Priors	$\sim 5\times$	$\sim 5\times$	0	**65.4**	**58.3**	**62.9**	**57.4**	**86.0**	**107.9**	**32.1**	**63.2**
w/o $\mathcal{L}_{alignment}$	$1\times$	$1\times$	-	64.0	57.2	61.3	55.9	83.4	106.3	30.7	61.1
ViPS (Ours)	$1\times$	$1\times$	0.252	64.6	57.6	62.0	56.5	85.5	**107.9**	31.6	62.5

within the same dataset (e.g., VSI-Bench). This observation aligns with our motivation that different foundation models exhibit distinct specializations, which shows that our framework successfully achieves context-aware prior selection by dynamically adjusting the fusion weights based on the specific scenario rather than statically relying on a single expert.

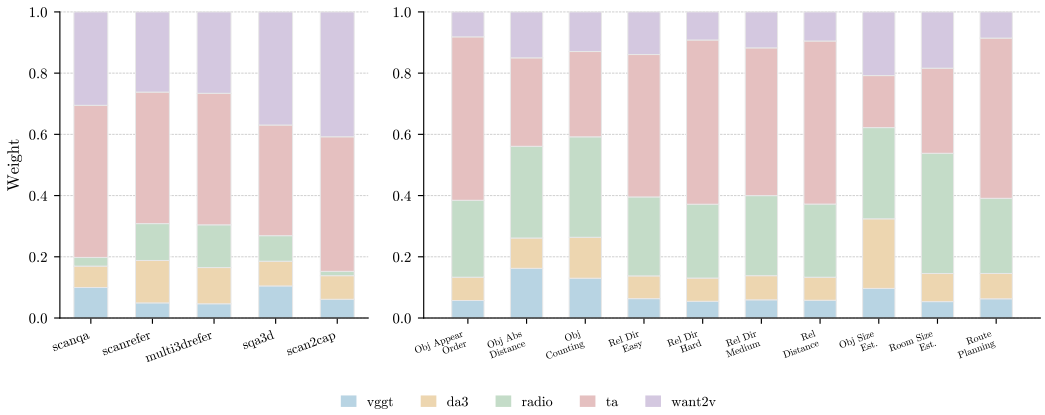


Figure 4: **Distribution of Dynamic Prior Weights.** This figure illustrates the distribution of the learned fusion weights (w_k in Equation 7) assigned to different foundation models. **Left:** The weight distribution on the test sets of ScanQA, ScanRefer, Multi3DRefer, SQA3D, and Scan2Cap. **Right:** The weight distribution across different question types in VSI-Bench.

4 Related Work

4.1 Spatial Understanding with Large Language Models

Spatial understanding, a foundational pillar for real-world interaction and reasoning, has witnessed a paradigm shift with the advent of LLMs [Brown et al., 2020, Ouyang et al., 2022, Touvron et al., 2023, Achiam et al., 2023]. Early attempts, such as PointLLM [Xu et al., 2024], PointBind [Guo et al.,

2023], GPT4Point [Qi et al., 2024], MiniGPT-3D [Tang et al., 2024], and Chat-3D [Wang et al., 2023], focused on aligning 3D point-cloud encoders directly with the LLM embedding space. To facilitate more effective cross-modal feature fusion, subsequent frameworks like Grounded-3D-LLM [Chen et al., 2024b], LL3DA [Chen et al., 2024a], 3D-LLaVA [Deng et al., 2025], and Inst3D-LLM [Yu et al., 2025] introduced advanced representation learning schemes. However, the inherent scarcity and noise of 3D point-cloud data often limit the scalability of these methods.

Recent research has gravitated towards video-based inputs. Prominent works such as 3D-LLM [Hong et al., 2023], Scene-LLM [Fu et al., 2024], Video-3D LLM [Zheng et al., 2025b], GPT4Scene [Qi et al., 2025] and SpatialStack [Zhang et al., 2026] establish dense correlations between 2D features and 3D scenes by building upon powerful pre-trained MLLMs [Li et al., 2024, Wang et al., 2024]. Specifically, Scene-LLM [Fu et al., 2024] captures fine-grained 3D knowledge through efficient 3D visual representation learning, while Video-3D LLM [Zheng et al., 2025b] introduces position-aware encodings for video sequences. Similarly, LLaVA-3D [Zhu et al., 2024a] achieves robust perception by learning a set of 3D voxels. While our work also falls within the video-input MLLM paradigm, we diverge from these approaches by exploring how to synergistically integrate prior knowledge from diverse foundation models to further elevate 3D spatial understanding.

4.2 Integration of Foundation Model Prior in MLLMs

The rapid evolution of foundation models—such as VGGT [Wang et al., 2025b], DepthAnything3 [Lin et al., 2025], and WAN [Wan et al., 2025]—has inspired a new line of research that injects diverse foundational priors into MLLMs to bolster their perceptual capabilities. Recent methods including VG-LLM [Zheng et al., 2025a], 3DRS [Huang et al., 2025], MiLO [Cao et al., 2025], VLM-3R [Fan et al., 2025], Vega-3D [Wu et al., 2026], ROSS3D [Wang et al., 2025a], and GeoThinker [Li et al., 2026] have demonstrated the efficacy of this paradigm. Specifically, VG-LLM employs a 3D visual geometry encoder to extract geometric priors from video sequences, while 3DRS aligns the latent space of MLLMs with VGGT via knowledge distillation. GeoThinker introduces an active perception mechanism for MLLMs to retrieve necessary geometric features, and Vega-3D injects world knowledge from video generation models, predicated on the hypothesis that these models inherently capture the underlying dynamics of scene transitions. In summary, existing efforts have predominantly focused on: (i) more efficient mechanisms for utilizing prior knowledge, and (ii) evaluating the performance of different individual model priors. In contrast, we propose **ViPS**, an efficient multi-model prior framework. By adaptively harmonizing various priors from disparate sources, our method significantly improves MLLM performance in spatial understanding tasks.

5 Conclusion

In this paper, we have introduced **ViPS**, a multi-model prior framework that unifies heterogeneous foundation-model priors within a single MLLM for spatial understanding. Our empirical study first reveals that no single foundation model dominates across tasks: different models supply distinct and complementary spatial knowledge. Building on this insight, ViPS couples an *Efficient Prior Proxy*, which distills the knowledge of multiple foundation models into lightweight branches sharing a common backbone and thereby avoids the linear cost of independent forward passes, with a *Dynamic Prior Fusion* mechanism that adaptively re-weights the resulting priors according to the input query and injects them through zero-initialized convolutions for stable training. On VSI-Bench and five ScanNet-series benchmarks, ViPS delivers state-of-the-art results in spatial reasoning, 3D visual grounding, dense captioning, and embodied question answering, while retaining single-encoder-level inference cost. These results suggest that context-aware harmonization of complementary priors is a promising path toward spatially grounded multimodal reasoning.

Acknowledgements

This work is supported by Hong Kong Research Grants Council – General Research Fund (Grant No. 17213825), Hong Kong Innovation and Technology Commission – Innovation and Technology Fund (Grant No. ITS/488/24FP), and HKU Seed Fund for PI Research.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Meng Cao, Haokun Lin, Haoyuan Li, et al. Seeing through imagination: Learning scene geometry via implicit spatial world modeling. *arXiv preprint arXiv:2512.01821*, 2025.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020.

Sijin Chen, Xin Chen, Chi Zhang, et al. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *CVPR*, 2024a.

Yilun Chen, Shuai Yang, Haifeng Huang, et al. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024b.

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Longvila: Scaling long-context visual language models for long videos. In *ICLR*, 2024c.

Zhe Chen, Jiannan Wu, Wenhai Wang, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024d.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, 2021.

Jiajun Deng, Tianyu He, Li Jiang, et al. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *CVPR*, 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Rao Fu, Jingyu Liu, Xilun Chen, et al. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, 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.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, et al. Radiov2.5: Improved baselines for agglomerative vision foundation models. In *CVPR*, 2025.

Yining Hong, Haoyu Zhen, Peihao Chen, et al. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Haifeng Huang, Zehan Wang, Rongjie Huang, et al. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bo Li, Yuanhan Zhang, Dong Guo, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Haoyuan Li, Qihang Cao, Tao Tang, et al. Thinking with geometry: Active geometry integration for spatial reasoning. *arXiv preprint arXiv:2602.06037*, 2026.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, et al. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*, 2025.

Haotong Lin, Sili Chen, Junhao Liew, et al. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, et al. Llava-next: Improved reasoning, ocr, and world knowledge. *arXiv preprint*, 2024.

Xinhang Liu, Yuxi Xiao, Donny Y Chen, et al. Trace anything: Representing any video in 4d via trajectory fields. *arXiv preprint arXiv:2510.13802*, 2025a.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, et al. Nvila: Efficient frontier visual language models. In *CVPR*, 2025b.

Xiaojian Ma, Silong Yong, Zilong Zheng, et al. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

Kun Ouyang, Yuanxin Liu, Haoning Wu, et al. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Zhangyang Qi, Ye Fang, Zeyi Sun, et al. Gpt4point: A unified framework for point-language understanding and generation. In *CVPR*, 2024.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, et al. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

Yuan Tang, Xu Han, Xianzhi Li, et al. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *ACM MM*, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haochen Wang, Yucheng Zhao, Tiancai Wang, et al. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *ICCV*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, et al. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zehan Wang, Haifeng Huang, Yang Zhao, et al. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

Xianjin Wu, Dingkang Liang, Tianrui Feng, et al. Generation models know space: Unleashing implicit 3d priors for scene understanding. *arXiv preprint arXiv:2603.19235*, 2026.

Runsen Xu, Xiaolong Wang, Tai Wang, et al. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, et al. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

Hanxun Yu, Wentong Li, Song Wang, et al. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In *CVPR*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.

Jiang Zhang, Shijie Zhou, Bangya Liu, et al. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. *arXiv preprint arXiv:2603.27437*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

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

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*, 2025a.

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

Chenming Zhu, Tai Wang, Wenwei Zhang, et al. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024a.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, et al. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, 2023.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, et al. Unifying 3d vision-language understanding via promptable queries. In *ECCV*, 2024b.

Appendix

A Detailed Training Dataset Description

A.1 ScanNet-series Dataset

For the 3D spatial understanding tasks, our training protocol aligns with the configurations established by previous works such as Video-3D LLM and 3DRS, which aggregate data from five distinct benchmarks, yielding a combined training set of approximately 223K sample pairs. Specifically, the composition includes:

- **ScanRefer and Scan2Cap:** Each dataset contributes 36,665 QA instances.
- **Multi3DRefer:** This multi-target grounding dataset provides 43,838 data entries.
- **ScanQA:** We utilize 26,515 question-answering pairs from this dataset.
- **SQA3D:** Serving as the largest subset in our collection, SQA3D contributes 79,445 samples.

With the exception of SQA3D, which is sourced from 518 unique 3D environments, the remaining four datasets are constructed upon 562 unique scans from the ScanNet corpus.

A.2 VSI-Bench Dataset

For the spatial reasoning evaluation, we follow the dataset composition presented in VLM-3R to train our model on VSI-Bench. Our training subset comprises a total of 207,658 instruction-tuning QA pairs. The dataset is structurally diverse, encompassing various scene typologies and specific reasoning tasks. The detailed distribution is as follows:

- **ScanNet++ QA:** The majority of the data originates from the high-fidelity ScanNet++ corpus, accounting for 135,119 QA pairs.
- **ScanNet QA:** An additional 51,630 QA pairs are derived from ScanNet.
- **Absolute Distance Estimation:** 16,805 samples are dedicated to object absolute distance estimation tasks.
- **Route Planning:** The dataset also includes 4,104 instances specifically designed to evaluate and enhance the model's route planning and navigational reasoning capabilities.

A.3 Details of Section 2.1

To ensure computational efficiency during our exploration, here we train the models on a 10% subset of the original training datasets used in 3DRS [Huang et al., 2025] and VLM-3R [Fan et al., 2025]. We conduct our empirical assessment on the VSI-Bench [Yang et al., 2025] and ScanNet-series benchmarks including ScanRefer [Chen et al., 2020], Multi3DRefer [Zhang et al., 2023b], Scan2Cap [Chen et al., 2021], ScanQA [Azuma et al., 2022], and SQA3D [Ma et al., 2022]. We employ Qwen2-VL [Wang et al., 2024] as our base MLLM. For each evaluated variant, we extract features from each foundation model via its respective encoder and project them into the same hidden dimension as the MLLM using an MLP. Then inject them into the image tokens of the MLLM via addition. The evaluation metrics are consistent with those in the main text.

B Training Detail

A comprehensive summary of all detailed hyperparameters and configurations is provided in Table 6 and Table 7. To align the external visual priors with the internal representations, we downsample all features extracted by the prior models to match the resolution of the MLLM's image tokens.

The trainable parameters in our ViPS framework include the LLM backbone (updated via LoRA), the multimodal projector (`mm_projector`), the zero-initialized convolutions and MLPs responsible for generating dynamic weights within the Dynamic Prior Fusion, and all parameters within the Efficient

Prior Proxy. Under our hardware configuration, the training process takes approximately 24 hours for the ScanNet-series datasets and 30 hours for the VSI-Bench dataset.

Table 6: **Training Hyperparameters and Hardware Configuration.** Detailed settings used to train ViPS (qwen2-vl-7b) on the ScanNet-series benchmarks.

Configuration / Hyperparameter	**Value**
Model Initialization	
Base MLLM	`LLaVA-Video-7B-Qwen2`
Vision Tower	`google/siglip-so400m-patch14-384`
Training Settings	
Hardware	$8 \times$ 48GB GPUs
Distributed Strategy	DeepSpeed ZeRO Stage 2
Data Precision	`bfloat16` (bf16)
Gradient Checkpointing	True
Optimization Hyperparameters	
Training Epochs	1
Per-device Batch Size	1
Gradient Accumulation Steps	2
Total Effective Batch Size	16
Base Learning Rate	2×10^{-5}
Warmup Ratio	0.03
LoRA Configurations	
LoRA Rank	512
LoRA Alpha	1024
Prior Features Alignment	
Target Spatial Resolution	14×14
Downsampling Method	Bilinear

Table 7: **Training Hyperparameters and Hardware Configuration.** Detailed settings used to train ViPS (qwen3-vl-8b) on the ScanNet-series and VSI-Bench datasets.

Configuration / Hyperparameter	**Value**
Model Initialization	
Base MLLM	`Qwen/Qwen3-VL-8B-Instruct`
Vision Tower	`Qwen3VLVisionModel`
Training Settings	
Hardware	$8 \times$ 48GB GPUs
Distributed Strategy	DeepSpeed ZeRO Stage 2
Data Precision	`bfloat16` (bf16)
Gradient Checkpointing	True
Optimization Hyperparameters	
Training Epochs	1
Per-device Batch Size	1
Gradient Accumulation Steps	2
Total Effective Batch Size	16
Base Learning Rate	2×10^{-5}
Warmup Ratio	0.03
LoRA Configurations	
LoRA Rank	256
LoRA Alpha	256
Prior Features Alignment	
Target Spatial Resolution	15×20
Downsampling Method	Bilinear

C Qualitative Analysis of Visual Priors

We conduct a prior analysis on the diverse foundation models evaluated in Section 2.1. As illustrated in Figure 5 (Left), the visualization demonstrates that priors extracted from distinct foundation models cluster into disparate regions within the latent space. Furthermore, the activation heatmaps in Figure 5 (Right) reveal that different models focus on varying structural and semantic cues within the identical scene. These results validate that different foundation models indeed encapsulate distinct representations and spatial cues due to their diverse training paradigms.

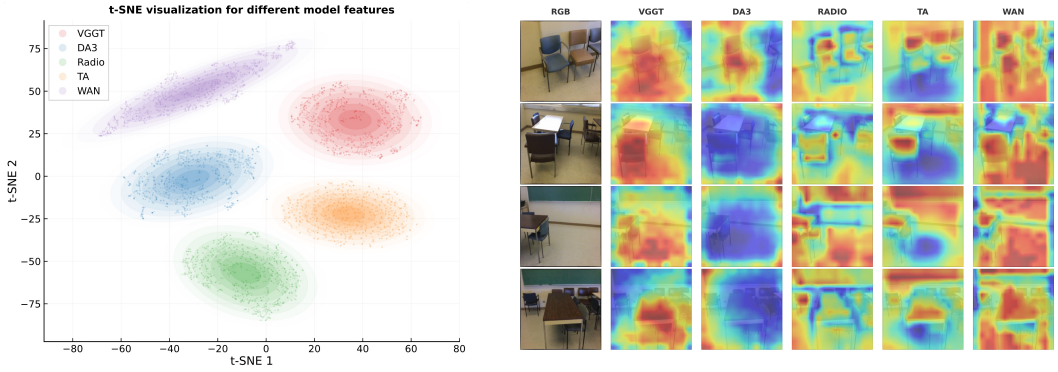


Figure 5: **Prior Analysis of Diverse Foundation Models. Left:** The t-SNE of visual prior features extracted from different foundation models. **Right:** Spatial heatmaps for different models

D Impact of Different Base Model

In the main paper, we default to using VGGT as the base foundation model within our Efficient Prior Proxy. To further investigate the robustness and flexibility of our ViPS framework, we conduct additional experiments by substituting VGGT with other foundation models as the base model, while keeping the rest of the framework unchanged.

As shown in Table 8, using different base models yields comparable and consistently strong performance across the ScanNet-series benchmarks. Although VGGT yields the best overall performance, we observe that TraceAnything achieves better results on certain metrics (e.g., 62.02 F1@0.25 on Multi3DRefer and 85.53 C@0.5 on Scan2Cap). Furthermore, all other foundation models maintain highly comparable performance when used as the base model. These results demonstrate that our Efficient Prior Proxy and Dynamic Prior Fusion mechanisms are robust and not strictly dependent on a specific base model choice.

Table 8: **Robustness to the Choice of Base Model.** Substituting VGGT with other foundation models in the Efficient Prior Proxy yields comparable performance across the ScanNet-series benchmarks, indicating that ViPS is not strictly dependent on a specific base-model choice.

Base Model	ScanRefer		Multi3DRefer		Scan2Cap	ScanQA		SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	C@0.5	CIDER	EM	EM
Baseline	62.1	54.6	59.6	54.4	81.4	104.6	30.5	60.8
VGGT	**64.6**	**57.6**	62.0	**56.5**	85.5	**107.9**	**31.6**	**62.5**
Wan	64.4	57.4	62.0	56.3	83.0	106.0	31.1	61.4
TraceAnything	63.8	56.7	**62.0**	56.4	**85.5**	107.0	31.3	61.8
DepthAnything	63.7	56.8	61.8	56.4	82.2	107.0	31.1	61.7
RADIO	64.1	57.2	61.7	56.4	82.6	107.3	31.2	62.5

E Visualization Result

To further demonstrate the effectiveness of our proposed ViPS framework, we provide visualization results in Figure 6. In this comparison, the baseline model is obtained by fine-tuning the `Qwen2-VL` architecture on the same training dataset used for our model. As illustrated, by synergistically harmonizing diverse visual priors, our ViPS framework achieves a more precise understanding of complex spatial relationships and generates more accurate responses compared to the baseline.

F Broader Impact

Positive Impacts. The capabilities introduced by our ViPS framework significantly advance the spatial reasoning and 3D spatial understanding of MLLMs. By effectively harmonizing diverse visual priors, our approach paves the way for more intelligent embodied agents, such as domestic robots and autonomous navigation systems. Furthermore, our Efficient Prior Proxy minimizes inference overhead, facilitating the deployment of sophisticated 3D perception systems on resource-constrained edge devices. This efficiency can democratize access to advanced AI, fostering innovations in assistive applications for visually impaired individuals to navigate complex environments safely.

Potential Negative Impacts and Mitigation. Endowing AI systems with highly accurate 3D spatial understanding capabilities inherently presents privacy and security risks. If deployed irresponsibly, such technologies could be exploited for unauthorized surveillance or the invasive reconstruction of private indoor spaces. To mitigate these ethical concerns, future real-world applications must incorporate robust anonymization pipelines—such as the automatic obfuscation of human faces and sensitive documents—prior to processing scene data. Additionally, strict access controls, transparent user consent mechanisms, and compliance with data protection regulations are essential to safeguard individual privacy.

G Limitation and Future Work

While ViPS demonstrates state-of-the-art performance in 3D spatial understanding, several limitations remain. First, due to computational resource constraints, our experiments are currently restricted to a 7B-parameter base MLLM and training datasets on the scale of a few hundred thousand samples. Scaling up both the model size and data volume could further enhance performance. Second, extracting spatial awareness from external visual priors is ultimately a sub-optimal workaround. A more fundamental future direction is to internalize this geometric knowledge directly during the MLLM pre-training phase, which could be achieved by introducing more diverse spatial datasets, fine-grained 3D sub-tasks, and tailored loss functions. Finally, our evaluations are currently confined to standard offline benchmarks. Deploying and validating spatial-aware MLLMs in real-world robotics and physical embodied AI scenarios remains an exciting avenue for future exploration.

Question_type: obj_appearance_order



Question: What will be the first-time appearance order of the following categories in the video: blanket, door, basket, table?
Options: A. door, table, basket, blanket B. door, blanket, basket, table
C. blanket, door, basket, table D. door, basket, blanket, table
GT: D Our Prediction:D Baseline Prediction:A

Question_type: route_planning



Question: You are a robot beginning at the tv facing the bed. You want to navigate to the trash bin. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the cabinet 3. [please fill in] 4. Go forward until the trash bin is on your right. You have reached the final destination.
Options: A. Turn Left, Turn Left B. Turn Right, Turn Left
C. Turn Back, Turn Left D. Turn Right, Turn Right
GT: B Our Prediction:B Baseline Prediction:D

Question_type: object_rel_distance



Question: Measuring from the closest point of each object, which of these objects (power strip, computer tower, bookshelf, laptop) is the closest to the whiteboard?
Options: A. power strip B. computer tower
C. bookshelf D. laptop
GT: C Our Prediction:C Baseline Prediction:D

Question_type: object_abs_distance



Question: Measuring from the closest point of each object, what is the distance between the tv and the stove (in meters)?
GT: 2.9 Our Prediction:2.9 Baseline Prediction:2.4

Figure 6: **Qualitative Visualization on VSI-Bench.** We compare our ViPS framework against the baseline, which is obtained by fine-tuning `Qwen2-VL` on our identical training dataset, showing that ViPS produces more accurate responses to complex spatial questions.

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: Yes, the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Yes, the paper discusses the limitations in Appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: No, the paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We believe all necessary details for reproduction are provided. Should there be any unintentional omissions regarding specific hyperparameters, we are fully committed to providing them upon request.

   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: [Yes]

Justification: Our code is publicly available at https://github.com/Visual-AI/ViPS, with instructions for environment setup, data preparation, Qwen3-VL training, and VSI-Bench evaluation.

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: Yes, the paper specifies all the training and test details.

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: Due to the significant computational resources required for MLLM training and inference, we report results from single runs. However, the performance gains are consistent across different benchmarks.

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: Yes, the paper provides sufficient information on the computer resources needed to reproduce the experiments.

   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: Yes, the paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss both potential positive societal impacts and negative societal impacts in appendix.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: No. We believe the paper does not pose any risks to society.

    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: Yes, the paper properly credits the creators or original owners of the assets used in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Yes.

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

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: This paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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