SpaR3D-MoE: Adaptive 3D Spatial Reasoning from Sparse Views Meets Geometry-Inductive Mixture-of-Experts

Haida Feng[1,2], Hao Wei[1,3†], Haolin Wang[1,2], Shiwei Li[1,2], Chade Li[1,2], and Yihong Wu[1,2,3†]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3] YUKUN Intelligent World, Beijing, China
{fenghaida2024@ia, weihao2019@ia, wanghaolin2023@ia, lishiwei2023@ia, lichade2021@ia, yhwu@nlpr.ia}.ac.cn

Abstract. Recent Multimodal Large Language Models (MLLMs) struggle to bridge the representational gap between 2D semantic understanding and 3D spatial geometry. Existing 3D-aware models either rely on costly 3D-specific data or utilize RGB-only inputs with heuristic sampling and monolithic, shallow fusion, which respectively disrupt essential spatiotemporal connectivity and induce modality contention across diverse spatial tasks. To overcome these bottlenecks, we introduce SpaR3D-MoE, an end-to-end framework that enables adaptive spatial reasoning by equipping MLLMs with geometry-aware capabilities from only sparse RGB inputs. First, we propose an adaptive spatiotemporal manifold sampling mechanism that constructs a geometry-aware spatiotemporal graph to extract informative keyframes, effectively mitigating sequence redundancy while preserving the scene's topological connectivity. Second, we introduce the heterogeneous geometry-inductive Mixture-of-Experts driven by an instruction-pose aware router, which adaptively routes multimodal tokens to specialized experts, resolving the cross-modal contention inherent in monolithic fusion. Extensive experiments on VSI-Bench, ScanQA, and SQA3D demonstrate that our method achieves state-of-the-art performance. Notably, SpaR3D-MoE achieves the highest average score of 63.5 on VSI-Bench, outperforming the strongest baseline by 7.8 absolute points, alongside relative improvements of 35.4% and 51.4% in Route Plan and Relative Direction tasks, respectively.

Keywords: 3D Spatial Reasoning · Mixture-of-Experts · Multimodal Large Language Models

1 Introduction

Three-dimensional (3D) spatial reasoning is a cornerstone of Embodied AI, enabling agents to understand complex environments, reason about spatial rela-

† Corresponding authors

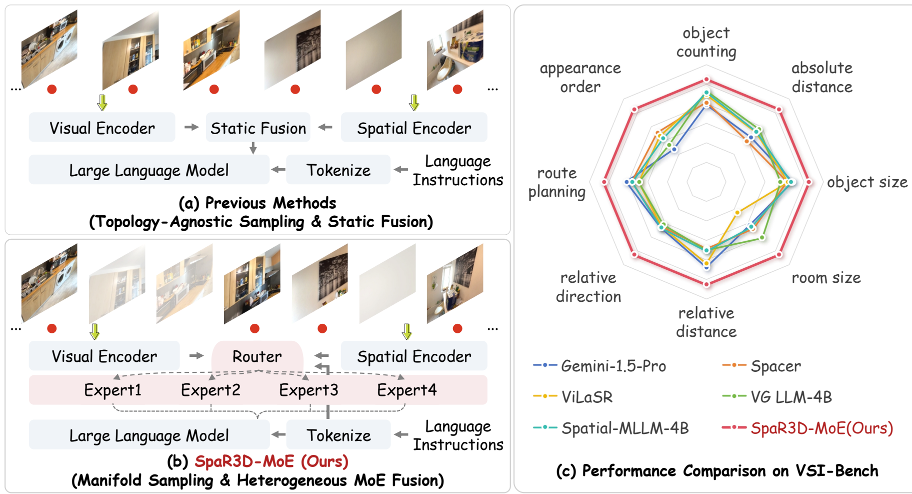

Fig. 1: Comparison of 3D spatial reasoning paradigms. Existing methods use topology-agnostic sampling and static monolithic fusion **(a)**. SpaR3D-MoE selects sparse keyframes via manifold-based sampling and routes multimodal tokens to specialized experts **(b)**, achieving strong performance across VSI-Bench spatial tasks **(c)**.

tionships, and ground natural language instructions in real-world contexts [2,47]. While Multimodal Large Language Models (MLLMs) [1, 4, 5, 10, 17, 23, 28] excel at 2D image and video understanding, extending them to the 3D physical world is hindered by a fundamental representational gap. Current models struggle with complex tasks like estimating metric distances or navigating spatial gaps, primarily because bridging 2D visual semantics with geometric 3D real-world alignments remains a challenge.

To bridge this representational gap, recent efforts have diverged into two paradigms aiming to map 3D geometric features into the MLLM latent space, either through explicit 3D structures or implicit visual representations. The first paradigm explicitly integrates 3D structures (e.g., point clouds, depth maps, or reconstructed meshes) directly into Large Language Models (LLMs) [7,8,13–16, 34, 47, 48]. These approaches typically lift multi-view RGB-D inputs or reconstructed scenes into 3D point clouds [14], then employ specialized 3D geometry encoders (e.g., PointNet++ [27]) to extract geometric features and project them into the textual latent space. While these explicit geometric priors significantly benefit physical grounding, this pipeline is fundamentally constrained by its rigid dependence on acquiring explicit 3D geometric proxies. Moreover, the intrinsic sparsity of point clouds often leads to the loss of rich visual details, compromising the fine-grained semantic understanding crucial for comprehensive scene reasoning. Conversely, the second paradigm focuses on scalable, 3D-aware MLLMs that rely solely on RGB sequences [35, 45]. These methods typically adopt a dual-encoder architecture, using a 2D visual encoder to extract semantic features and a spatial encoder, often initialized from visual geometry foundation models [33], to recover implicit 3D structural features from 2D RGB inputs. Despite bypassing costly 3D-specific data for easier applicability, this paradigm remains constrained by heuristic sampling and static, monolithic fusion. Regarding frame sampling, current methods rely on topology-agnostic selection strategies, such as rigid uniform sampling or discrete voxel maximization heuristics. By treating frames as isolated viewpoints, these approaches introduce spatiotemporal redundancy and overlook the intrinsic spatiotemporal manifold of the 3D scene, potentially missing critical spatial frames. At the multimodal integration level, monolithic fusion indiscriminately projects heterogeneous semantic textures and geometric structures into a shared latent space. Such architectural inflexibility induces cross-modal contention when faced with diverse task requirements, thereby hindering fine-grained spatial reasoning, as illustrated in Fig. 1(a).

To address these limitations, we introduce SpaR3D-MoE, a framework that shifts the paradigm toward adaptive spatial reasoning from sparse RGB inputs, as illustrated in Fig. 1(b). Specifically, we propose an **A**daptive **S**patiotemporal **M**anifold **S**ampling (ASMS) mechanism to extract informative keyframes. By constructing a spatiotemporal graph from viewpoint-dependent geometry and camera ego-motion, regulated by a motion-aware quality gate, it filters redundancy while preserving essential spatiotemporal connectivity of the scene. Furthermore, to mitigate the cross-modal contention in monolithic fusion, we present a **H**eterogeneous **G**eometry-**I**nductive Mixture-of-Experts (HGI-MoE). To the best of our knowledge, this is the first work that introduces MoE architectures into 3D spatial reasoning. Inspired by the functional specialization of the human brain [32], this module is driven by an **I**nstruction-**P**ose **A**ware **R**outer (IPAR), which adaptively routes multimodal features to specialized experts based on linguistic intent and camera ego-motion. Crucially, these experts exhibit emergent specialization, performing varying levels of cross-modal fusion tailored to meet diverse spatial reasoning requirements. This dynamic dispatching effectively mitigates cross-modal interference by establishing disentangled reasoning pathways. To ensure stable expert convergence, we also introduce a load-balancing loss function that regularizes the expert assignment and prevents routing collapse. Extensive experiments on large-scale benchmarks, including VSI-Bench [41], ScanQA [3], and SQA3D [25], demonstrate that SpaR3D-MoE achieves state-of-the-art (SOTA) performance, validating its adaptive reasoning across diverse challenging 3D spatial reasoning tasks from sparse RGB-only views.

The main contributions of SpaR3D-MoE are outlined as follows:

- We propose ASMS mechanism, which constructs a spatiotemporal graph with a motion-aware quality gate to adaptively extract informative keyframes, reducing redundancy while preserving manifold connectivity, achieves a 10.6% improvement over uniform sampling on the VSI-Bench Route Plan task.
- We propose HGI-MoE, which introduces the Mixture-of-Experts paradigm to 3D spatial reasoning, where an IPAR adaptively dispatches multimodal features to specialized experts for varying levels of fusion, mitigating modality contention and robustly fulfilling diverse spatial reasoning demands.

– Empowered by these designs, we introduce SpaR3D-MoE, an end-to-end adaptive framework that endows MLLMs with physically-grounded spatial intelligence from sparse RGB-only views. It achieves SOTA across diverse benchmarks, notably surpassing the strongest VSI-Bench baseline by 7.8 absolute points, while delivering remarkable relative gains of 35.4% and 51.4% in complex navigation and fine-grained metric estimation, respectively.

2 Related Work

2.1 Multimodal Large Language Models

Recent Multimodal Large Language Models (MLLMs) [4, 5, 17, 28, 31, 44] have achieved impressive capabilities in image and video understanding. These models typically integrate visual encoders with large language models to process and generate text based on visual inputs. They have been applied in various domains, including visual question answering and multimodal dialogue systems. However, while current MLLMs can reason about complex relationships from images and videos, they struggle to ground these relations in the 3D physical world. Primarily trained on 2D image-text pairs that prioritize semantic content over geometric structure, these models lack a necessary understanding of 3D spatial relationships and geometry. Consequently, with standard visual encoders compressing spatial details for semantic alignment [6,30], 2D-based MLLMs often hallucinate when faced with complex 3D spatial understanding tasks.

2.2 3D-Aware Multimodal Large Language Models

Recent research increasingly leverages pre-trained MLLMs to tackle complex 3D scene understanding and reasoning tasks. Existing methods can be broadly grouped by how they introduce spatial information. Early works [7, 8, 13–16, 34, 47, 48] mainly rely on 2.5D or 3D representations, such as posed RGB-D data, point clouds, or voxel grids. Although these explicit geometric inputs provide strong 3D grounding, their dependence on depth sensors or reconstructed 3D assets limits scalability in RGB-only scenarios. Recent RGB-based methods [35,45] alleviate this issue by using VGGT [33] to extract implicit 3D geometric features from monocular inputs and align them with MLLMs. Other studies improve spatial reasoning from complementary perspectives, such as reinforcement learning for video spatial reasoning [26] and structured 2D representations for perception-guided reasoning [49]. Despite these advances, preserving sparse-view spatiotemporal topology while adaptively fusing visual and geometric features for diverse spatial tasks remains insufficiently explored. In contrast, our ASMS adaptively samples sparse keyframes to preserve essential spatiotemporal connectivity, while HGI-MoE dynamically activates specialized experts for task-aware visual-geometric fusion, thereby enhancing 3D spatial reasoning performance.

2.3 Mixture-of-Experts Framework

The MoE paradigm has fundamentally transformed the scaling of LLMs by decoupling model capacity from inference latency [18,29,37]. By dynamically routing tokens to a sparse subset of active parameters, these architectures achieve massive representational power without imposing prohibitive computational bottlenecks. Building on this foundation, recent foundation models [11,38–40] have further refined sparse routing mechanisms to achieve unprecedented scale and efficiency in pure language modeling. Beyond text, this paradigm has been extended to the multimodal domain [20,21], where modality-specific encoders and multi-stage tuning successfully scale diverse inputs without sparsity degradation. Motivated by the efficacy of MoE in modeling heterogeneous data, we introduce a multimodal MoE framework tailored for various complex 3D spatial reasoning tasks. To accommodate the diverse distributions of data inputs and instructional intents, our framework dynamically routes features to heterogeneous experts, enabling efficient and robust scene comprehension.

3 Methodology

3.1 Problem Formulation and Framework

Let $\mathcal{V} = \{v_i\}_{i=1}^{N_k}$ denote a continuous sequence of embodied visual observations capturing a complex 3D environment, and $\mathcal{Q}$ represent a natural language instruction specifying a spatial reasoning task. Our primary objective is to autoregressively decode a precise textual response $\mathcal{A} = \{a_t\}_{t=1}^{T}$.

Processing RGB video sequences for spatial understanding through topology-agnostic sampling reduces spatiotemporal redundancy but disrupts key topological connectivity, while the subsequent shallow and monolithic feature fusion inevitably leads to cross-modal contention. To overcome these limitations, we decouple the unified spatial reasoning task into two synergistic sub-problems. First, we formalize keyframe extraction as an information maximization problem on the spatiotemporal manifold under a sparsity constraint $N_n \ll N_k$. From the optimal sparse subset $\mathcal{V}_s \subset \mathcal{V}$, we extract their 2D visual features $\mathbf{F}_{\mathrm{2D}}$, 3D geometric features $\mathbf{F}_{\mathrm{3D}}$, and corresponding camera poses $\mathcal{P}$. Subsequently, we formalize multimodal alignment as a conditional routing policy. We introduce a routing function Φ_{MoE} that adaptively fuses these heterogeneous features, guided by the language instruction $\mathcal{Q}$ and spatial context from camera poses $\mathcal{P}$. The unified objective is to maximize the conditional likelihood of the target sequence:

$$\mathcal{A}^* = \arg\max_{\mathcal{A}} \prod_{t=1}^{T} p_\theta\Big(a_t \mid a_{<t}, \Phi_{\mathrm{MoE}}\big(\mathbf{F}_{\mathrm{2D}}, \mathbf{F}_{\mathrm{3D}} \mid \mathcal{P}, \mathcal{Q}\big), \mathcal{Q}\Big), \tag{1}$$

where p_θ denotes the probability distribution parameterized by the generative language decoder $\mathcal{F}_{\mathrm{LLM}}$ with learnable parameters θ.

Our end-to-end SpaR3D-MoE framework is shown in Fig. 2. First, long videos are sampled into sparse keyframes utilizing ASMS (Sec. 3.2). Next, visual and

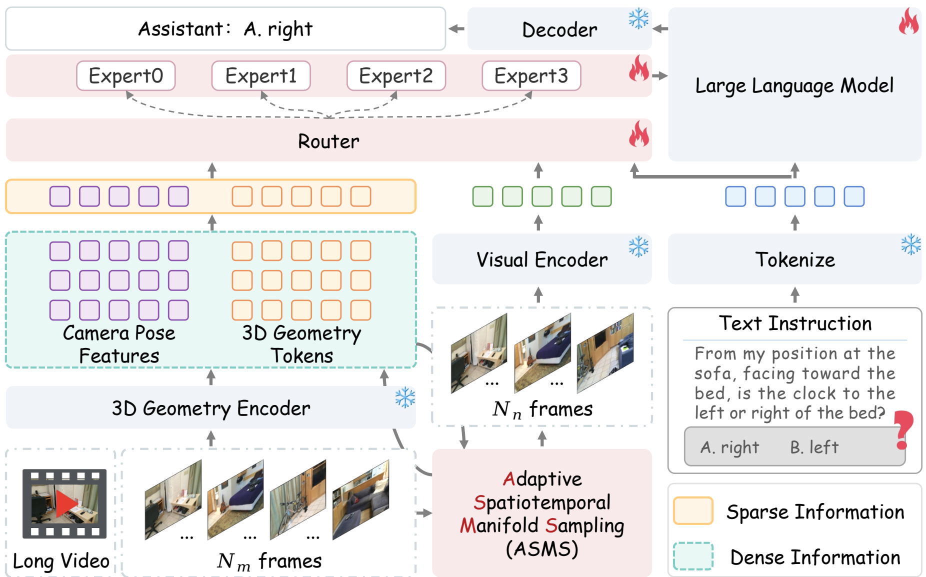

Fig. 2: Overview of SpaR3D-MoE framework. Given long videos, the ASMS constructs a spatiotemporal graph to adaptively distill N_m candidates into N_n informative keyframes. Along with text instructions, multimodal features are dynamically routed via the instruction-pose aware router to specialized experts ($E_0 - E_3$), then aggregated to drive the MLLM for spatial reasoning.

instruction tokens are encoded by Qwen3-VL [4], while 3D geometric and pose features are extracted by VGGT [33]. These features are then dynamically dispatched to four geometry-inductive experts via IPAR (Sec. 3.3). Finally, the adaptive multimodal features are projected into the LLM for autoregressive spatial reasoning, jointly optimized with a routing load-balancing penalty (Sec. 3.4).

3.2 Adaptive Spatiotemporal Manifold Sampling

Continuous embodied observations naturally reside on a low-dimensional data manifold embedded within a high-dimensional multimodal feature space. To process these continuous streams with bounded computational overhead, we first uniformly downsample the N_k-frame video into $N_m = 128$ candidates, preserving macroscopic scene topology. Since further rigid uniform sampling to achieve extreme sparsity ($N_n \ll N_m$, e.g., $N_n \in 8, 16, 32$) disrupts this intrinsic connectivity, we introduce the ASMS mechanism, which extracts N_n topologically crucial keyframes by approximating the manifold via a spatiotemporal graph.

To begin with, we establish a composite distance metric $\mathcal{D}(i, j)$ to quantify the frame-to-frame relations as the edge weights of our spatiotemporal graph. This metric integrates the relative camera displacement, implicit geometric dissimilarity, and temporal distance across any two frames:

$$\mathcal{D}(i,j) = \bar{\mathcal{D}}_{trans}(i,j) + \gamma\bar{\mathcal{D}}_{geo}(i,j) + \omega\bar{\mathcal{D}}_{tmp}(i,j), \tag{2}$$

where $\bar{\mathcal{D}}_{trans}$ is the normalized L_2 distance of 3D translations from 6D poses predicted by VGGT [33], serving as a spatial anchor to ensure broad trajectory coverage. To account for changes in viewing direction, $\bar{\mathcal{D}}_{geo}$ captures rotation-induced variations using cosine similarity of implicit 3D features encoded by VGGT, preventing redundant sampling at the same position. Meanwhile, $\bar{\mathcal{D}}_{tmp}$ uses normalized temporal interval to preserve temporal progression. We set $\gamma = 0.5$ and $\omega = 0.6$ to balance geometric diversity and temporal stability.

Building upon the distance metric, we introduce a node-level quality score $\mathcal{S}_i$ to ensure the informational validity of the sampled manifold. Relying solely on spatial distances risks anchoring the graph on uninformative regions, such as textureless blank walls. To address this, we quantify the visual richness of each candidate frame by computing the average L_2 norm of its Top-K_v patch tokens $\mathbf{f}_{i,k}$. This prevents sparse foreground features from being diluted by massive background tokens. Additionally, to suppress motion blur caused by fast camera movement, we penalize this score with an exponential decay factor based on the instantaneous camera velocity v_i. The final quality score is defined as:

$$\mathcal{S}_i = \left(\frac{1}{K_v} \sum_{k \in \mathcal{T}_{K_v}(i)} \|\mathbf{f}_{i,k}\|_2 \right) \cdot \exp\left(-\frac{v_i}{\bar{v}}\right) + \epsilon, \tag{3}$$

where $\mathcal{T}_{K_v}(i)$ is the index set of the Top-K_v tokens for frame i, $\bar{v}$ is the mean trajectory velocity, and ϵ ensures a baseline sampling probability.

Ultimately, to sparsify the dense video while preserving its underlying manifold structure, we formulate keyframe extraction as quality-gated farthest point sampling (FPS) process on the approximated spatiotemporal graph. At each iteration, we greedily select the candidate frame i^* that maximizes its shortest distance to the already sampled subset $\mathcal{K}$, modulated by its quality score:

$$i^* = \arg\max_{i \notin \mathcal{K}} \Big[\big(\min_{j \in \mathcal{K}} \mathcal{D}(i,j) \big) \cdot (1 + \lambda \mathcal{S}_i) \cdot \mathcal{M}(i) \Big], \tag{4}$$

where $\lambda = 3.0$ balances structural coverage and visual richness. To ensure topological robustness, we formulate $\mathcal{M}(i)$ as a motion-aware quality gate. Nodes scoring below $\tau = 0.6\bar{\mathcal{S}}$ are penalized via $\mathcal{M}(i) = 0.1$, while reliable nodes retain $\mathcal{M}(i) = 1.0$. This effectively filters low-quality frames, yielding a sparse yet informative representation that covers the spatiotemporal manifold.

3.3 Heterogeneous Geometry-Inductive Mixture-of-Experts

Instruction-Pose Aware Router. As illustrated in Fig. 3, our router jointly leverages the language instruction tokens q, 2D visual tokens v, 3D geometric tokens g, and camera pose features p. Specifically, the first three components $(q, v,$ and $g)$ are linearly projected and concatenated to capture task-specific

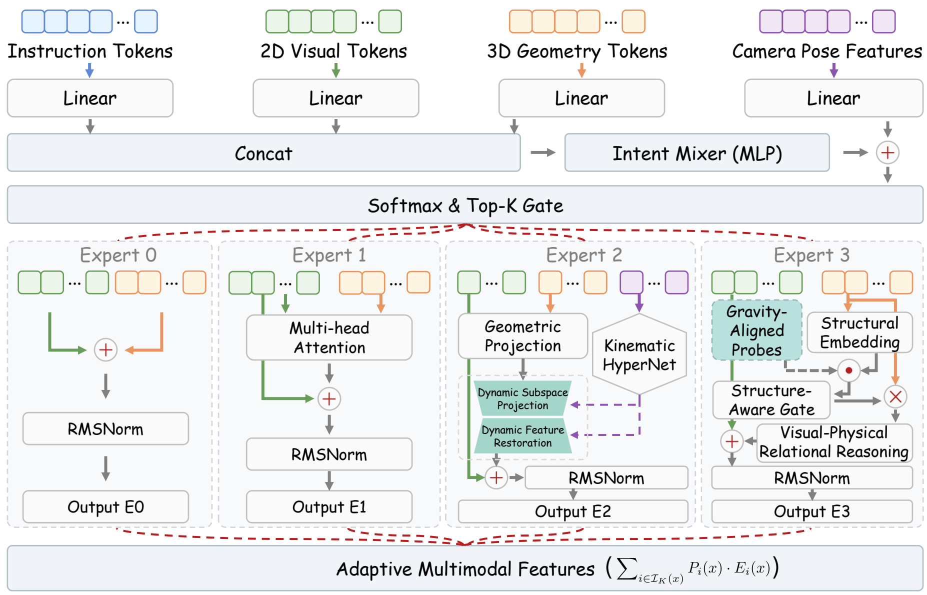

Fig. 3: Detailed architecture of the HGI-MoE. The router processes multimodal features to generate routing probabilities, dynamically dispatching tokens to the most suitable experts ($E_0 - E_3$), whose outputs are aggregated via a probability-weighted sum into adaptive multimodal features.

multimodal correlations via an Intent Mixer (MLP). Meanwhile, to handle viewpoint variations caused by camera motion, the camera pose feature p is linearly projected and added directly into the routing logit space as a global spatial bias. The routing logits $L \in \mathbb{R}^{N\times E}$ are computed as:

$$L = \mathrm{MLP}([\mathbf{W}_q q\,;\,\mathbf{W}_v v\,;\,\mathbf{W}_g g]) + \mathbf{W}_p p, \tag{5}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. This formulation ensures that the Top-K expert activation is dynamically steered by both the user instruction and the underlying 3D spatial context.

Heterogeneous Geometry-Inductive Experts. As illustrated in the lower section of Fig. 3, the routed multimodal tokens are dispatched to a specialized pool of four architecturally diversified experts. Each expert employs a tailored feature fusion mechanism to tackle distinct requirements of 3D spatial reasoning.

Holistic Representation Expert (E_0). While static shallow fusion often induces modal conflicts in monolithic architectures, we isolate this fundamental operation as a conditionally activated expert. E_0 achieves explicit modality integration via $E_0(v, g) = \mathrm{RMSNorm}(v + g)$. Retaining this primitive baseline serves a dual purpose. It provides a natural contrast to our other specialized spatial experts and explicitly highlights the adaptive superiority of our MoE design. Unlike static monolithic fusion, this simple additive branch is dynamically assigned to tokens that do not require complex spatial transformations.

Geometric-Semantic Cross-Attention Expert (E_1). To establish a precise mapping between 2D visual semantics and 3D geometry, this expert employs a multi-head cross-attention mechanism. By using the visual tokens v as queries and the geometric tokens g as keys and values, E_1 explicitly injects 3D spatial priors into the 2D visual representations. Incorporating a residual connection, the aggregated output is formulated as $E_1(v, g) = \text{RMSNorm}(v + \text{CrossAttn}(Q = v, K = g, V = g))$. This adaptive alignment tightly couples 2D appearance with 3D geometry information, equipping the model with fine-grained spatial awareness for complex scene reasoning.

Pose-Conditioned Dynamic Adapter Expert(E_2). To bridge the gap between static 3D geometry and large-baseline camera ego-motion across sparse views, E_2 operates as a pose-driven dynamic adapter. Rather than treating camera parameters as simple concatenated tokens, we leverage the view-specific pose features p to actively modulate geometric features g with view-dependent context. Using a HyperNet $\mathcal{H}$, the egocentric motion state is encoded into dynamic adaptation weights. Specifically, following an initial projection to obtain the base geometric embedding g_{emb}, these weights parametrize a low-rank bottleneck consisting of dynamic subspace projection and a subsequent feature restoration. The resulting pose-modulated output is then integrated into visual tokens v via a learnable α-scaled residual connection:

$$\begin{aligned} g_{adapted} &= \mathcal{F}_{\text{adapt}}(g_{emb} \mid \mathcal{H}(p)), \\ E_2(v, g, p) &= \text{RMSNorm}(v + \alpha \cdot \text{Dropout}(g_{adapted})), \end{aligned} \tag{6}$$

where $\mathcal{F}_{\text{adapt}}$ denotes the rank-constrained dynamic transformation. This conditionally modulated design effectively transforms viewpoint-invariant geometric priors into a motion-aware feature space tailored for sparse visual inputs.

Gravity-Aligned Structural Expert (E_3). To ground physical metrics within complex 3D environments, E_3 establishes canonical references by identifying orthogonal structural foundations. Specifically, we introduce N_p learnable structural probes $\mathbf{\Phi} \in \mathbb{R}^{N_p \times D}$ to capture gravity-aligned physical priors corresponding to the dominant scene layouts (e.g., floor planes). The geometric features g are first projected into a structural subspace g_{struct}. A structure-aware mask M is then derived from the cross-affinity between g_{struct} and $\mathbf{\Phi}$ to filter out unstructured spatial noise. These gated geometric features $\tilde{g}$ are subsequently integrated with the visual tokens v through a relational mapping network $\mathcal{F}_{\text{rel}}$:

$$\begin{aligned} M &= \sigma((g_{struct}\mathbf{\Phi}^T)\mathbf{W}_{attn}), \quad \tilde{g} = g \odot M, \\ E_3(v, g) &= \text{RMSNorm}(v + \mathcal{F}_{\text{rel}}([v; \tilde{g}])), \end{aligned} \tag{7}$$

where $\mathbf{W}_{attn}$ is a dimension-matching projection, and σ denotes the sigmoid activation. This design enables the model to encode global structural layouts, anchoring spatial reasoning in a consistent, gravity-aware coordinate system.

Adaptive Multimodal Aggregation. Given the routing logits $f(x)$, a Top-K gating mechanism computes the adaptive probability distribution across the expert pool for each token x. The final multimodal representation is aggregated as a probability-weighted sum of the activated experts' outputs:

$$P_i(x) = \frac{e^{f(x)_i}}{\sum_{j \in \mathcal{I}_K(x)} e^{f(x)_j}}, \quad \mathrm{MoE}(x) = \sum_{i \in \mathcal{I}_K(x)} P_i(x) \cdot E_i(x), \tag{8}$$

where $\mathcal{I}_K(x)$ denotes the index set of the K actively selected experts for token x. This adaptive activation ensures dynamic execution paths precisely tailored to the specific semantic queries and spatial complexities of the scene.

3.4 Optimization Objectives

We train our framework end-to-end with a composite objective:

$$\mathcal{L} = \mathcal{L}_{gen} + \lambda_{moe} \mathcal{L}_{moe}, \tag{9}$$

where $\mathcal{L}_{gen}$ is the standard cross-entropy loss for language generation, and λ_{moe} is a scaling factor for the auxiliary routing penalty. Due to the sparse activation mechanism of MoE, expert utilization can become imbalanced during training, potentially leading to expert collapse. To address this, we introduce a load-balancing loss $\mathcal{L}_{moe}$ to encourage uniform expert usage across the token sequence. The load-balancing objective is defined as:

$$\mathcal{L}_{moe} = N_e \sum_{i=1}^{N_e} \bar{P}_i \cdot \rho_i, \tag{10}$$

where N_e is the total number of experts, the mean routing probability $\bar{P}_i$ and activation frequency ρ_i for expert i across all N batch tokens are formulated as:

$$\bar{P}_i = \frac{1}{N} \sum_{t=1}^{N} P_i(x_t), \quad \rho_i = \frac{1}{N} \sum_{t=1}^{N} \mathbb{I}\left(i \in \mathcal{I}_K(x_t)\right), \tag{11}$$

where $P_i(x_t)$ is the routing probability of expert i for token x_t, $\mathcal{I}_K(x_t)$ denotes the K highest-scoring expert indices, and $\mathbb{I}(\cdot)$ denotes the indicator function.

4 Experiments

To comprehensively evaluate SpaR3D-MoE across diverse 3D spatial tasks, we benchmark the model on three datasets that encompass varying levels of spatial reasoning complexity: VSI-Bench [41] for general spatial reasoning, ScanQA [3] for fine-grained scene question answering, and SQA3D [25] for situated reasoning. Due to space constraints, implementation details are provided in Appendix.

Table 1: Comparison with SOTA methods on VSI-Bench. We evaluate the Qwen3VL models using the lmms-eval [42]. Notably, our SpaR3D-MoE achieves SOTA performance using solely 32 non-uniformly sampled sparse frames. Best results in each category are highlighted in **bold**, and the second-best are underlined.

Methods	Avg.	Numerical Answer Task				Multiple-Choice Answer Task			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Proprietary Models (API)									
GPT-4o [17]	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5-Flash [28]	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5-Pro [28]	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-source Models									
InternVL2-8B [10]	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B [10]	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
InternVL3-78B [50]	48.5	**71.2**	**53.7**	44.4	39.5	**55.9**	39.5	28.9	54.5
LongVILA-8B [9]	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5
VILA-1.5-40B [22]	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
LongVA-7B [43]	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7
LLaVA-NeXT-Video-7B [44]	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B [44]	40.9	48.9	22.8	57.4	35.3	42.4	36.7	**35.0**	48.6
LLaVA-OneVision-7B [19]	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B [19]	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
Qwen2.5VL-7B [5]	33.0	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8
Qwen3VL-4B [4]	54.8	66.4	43.4	**74.2**	60.0	53.0	46.2	32.0	63.1
Qwen3VL-8B [4]	**55.7**	68.0	45.8	73.5	**60.3**	53.7	**46.3**	32.5	**65.5**
Spatial-Aware MLLMs									
VG LLM-4B [45]	46.1	66.4	36.6	55.2	56.3	40.8	43.4	30.4	39.5
Spacer [26]	45.5	57.8	28.2	59.9	47.1	40.1	45.4	33.5	52.1
ViLaSR [36]	45.4	63.5	34.4	60.6	30.9	48.9	45.2	30.4	49.2
Spatial-MLLM [35]	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
SpaR3D-MoE (Ours)	**63.5**	**71.3**	**48.0**	**72.8**	**69.6**	**58.7**	**70.1**	**44.0**	**73.5**

4.1 Comparison with State-of-the-Art Methods

Evaluation on VSI-Bench. As summarized in Table 1, SpaR3D-MoE achieves a new SOTA with a 63.5 average on VSI-Bench, surpassing the strongest baseline Qwen3VL-8B by 7.8 points, while yielding relative gains of 35.4% in Route Plan and 51.4% in Relative Direction. Notably, our framework also exceeds the leading commercial model, Gemini-1.5-Pro, by a margin of 18.1 points. This performance gap is particularly evident in complex tasks, such as Route Plan (44.0 vs. 36.0) and Relative Direction (70.1 vs. 46.3). Moreover, SpaR3D-MoE achieves these advantages utilizing only 32 non-uniformly sampled sparse frames, in contrast to the dense input required by Gemini-1.5-Pro ($\sim$85 frames). We attribute these gains to our ASMS mechanism, which adaptively selects high-quality sparse keyframes from long video sequences to construct a comprehensive and informative scene context. Building upon this, our HGI-MoE architecture employs a dynamic router guided by instructions and camera poses to dispatch input tokens to dedicated experts. This mechanism ensures an adaptive and comprehensive fusion of multimodal features, driving performance improvements across diverse complex spatial reasoning tasks.

Evaluation on ScanQA. As detailed in Tab. 2, SpaR3D-MoE performs competitively on the ScanQA validation set, a benchmark that requires both semantic grounding and spatial reasoning. Our framework achieves SOTA performance among video-based models using sparse RGB inputs, reaching an EM@1

Table 2: Evaluations on the ScanQA (val). B-1 to B-4 are BLEU-n scores. Best results in each category are highlighted in **bold**, and the second-best are underlined.

Methods	ScanQA (val)							
	EM@1	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
Task-Specific Models								
ScanQA [3]	21.1	30.2	20.4	15.1	10.1	33.3	13.1	64.9
3D-Vista [51]	22.4	-	-	-	10.4	35.7	13.9	69.6
3D/2.5D-Input Models								
3D-LLM [8]	20.5	39.3	25.2	18.4	12.0	35.7	14.5	69.4
LL3DA [7]	-	-	-	-	13.5	37.3	15.9	76.8
Chat-Scene [15]	21.6	43.2	29.1	20.6	14.3	41.6	18.0	87.7
3D-LLaVA [12]	-	-	-	-	<u>17.1</u>	<u>43.1</u>	<u>18.4</u>	<u>92.6</u>
Video-3D LLM [46]	**30.1**	**47.1**	**31.7**	**22.8**	**16.2**	**49.0**	**19.8**	**102.1**
Video-Input Models								
Qwen2.5-VL-3B [5]	15.4	22.5	13.1	8.1	3.8	25.4	9.7	47.4
Qwen2.5-VL-7B [5]	19.0	27.8	13.6	6.3	3.0	29.3	11.4	53.9
Qwen2.5-VL-72B [5]	24.0	26.8	17.8	14.6	12.0	35.2	13.0	66.9
LLaVA-Video-7B [44]	-	39.7	26.6	9.3	3.1	44.6	17.7	88.7
Oryx-34B [24]	-	38.0	24.6	-	-	37.3	15.0	72.3
Spatial-MLLM [35]	<u>26.3</u>	<u>44.4</u>	<u>28.8</u>	<u>21.9</u>	<u>14.8</u>	<u>45.0</u>	<u>18.4</u>	<u>91.8</u>
SpaR3D-MoE (Ours)	**30.4**	**46.4**	**31.6**	**23.3**	**17.1**	**48.6**	**19.5**	**101.5**

of 30.4 and a CIDEr of 101.5. Specifically, it surpasses Video-3D LLM in Exact Match (30.4 vs. 30.1) while exceeding 3D-LLaVA by a significant margin in CIDEr (101.5 vs. 92.6). These results demonstrate that by dispatching multimodal features to suitable experts guided by instructions and camera poses, our approach seamlessly adapts to diverse task instructions, achieving robust spatial understanding from sparse RGB keyframes without costly 3D-specific data.

Evaluation on SQA3D. Beyond static scene understanding, we evaluate SpaR3D-MoE on the SQA3D benchmark to assess its situated reasoning capabilities. As reported in Tab. 3, our framework establishes a new SOTA among video-based methods with an average EM@1 of 58.3. Notably, it even outperforms the explicit 3D-based Video-3D LLM in fine-grained categories like *What* and *How*. We attribute this competitive performance to a synergistic design where our ASMS provides rich scene context while preserving topological connectivity, establishing an informative foundation. Furthermore, the HGI-MoE adaptively activates specialized experts guided by situated instructions and camera poses, effectively leveraging multimodal features to achieve precise and robust spatial understanding from RGB-only inputs.

4.2 Ablation Study

To validate the architectural designs of SpaR3D-MoE, we perform comprehensive ablation studies on VSI-Bench [41]. We systematically analyze the contributions of heterogeneous experts, the role of multimodal routing guidance, and the impact of our adaptive sampling mechanism across different frame densities.

Effectiveness of Heterogeneous Geometry-Inductive Experts. To evaluate expert specialization within HGI-MoE, we selectively mask individ-

Table 3: Evaluations on the SQA3D (test). All scores are reported in EM@1. Best results in each category are highlighted in **bold**, and the second-best are underlined.

Methods	SQA3D (test)						
	What	Is	How	Can	Which	Others	Avg.
Task-Specific Models							
SQA3D [25]	31.6	63.8	46.0	69.5	43.9	45.3	46.6
3D-Vista [51]	34.8	63.3	45.4	69.8	47.2	48.1	48.5
3D/2.5D-Input Models							
Scene-LLM [13]	40.9	69.1	45.0	**70.8**	47.2	52.3	54.2
Chat-Scene [15]	45.4	67.0	52.0	69.5	49.9	55.0	54.6
Video-3D LLM [46]	**51.1**	**72.4**	**55.5**	69.8	**51.3**	**56.0**	**58.6**
Video-Input Models							
Qwen2.5-VL-3B [5]	34.8	52.1	39.8	52.7	45.6	47.0	43.4
Qwen2.5-VL-7B [5]	39.7	56.6	41.1	55.9	47.6	47.2	46.5
Qwen2.5-VL-72B [5]	41.7	56.3	41.5	55.6	44.5	48.0	47.0
LLaVA-Video-7B [44]	42.7	56.3	47.5	55.3	50.1	47.2	48.5
Spatial-MLLM-4B [35]	45.9	71.6	55.1	**69.5**	52.0	53.0	55.9
SpaR3D-MoE (Ours)	**51.6**	**72.6**	**56.3**	69.2	**52.1**	**54.2**	**58.3**

ual experts during inference to analyze their performance impacts, as detailed in Tab. 4. First, masking the **holistic representation expert** (E_0) reduces overall performance by 1.7. Instead of uniformly affecting all tasks, it significantly impairs Relative Direction and Appearance Order by 4.5 and 2.6, respectively. This validates our motivation to isolate simple additive fusion. Without E_0, tokens not requiring complex spatial processing are routed through high-order experts, resulting in over-processing of raw features and diminishing the specialized capacity of other branches. Furthermore, masking the **geometric-semantic cross-attention expert** (E_1) specifically degrades fine-grained spatial relational tasks, dropping Relative Distance and Object Counting by 1.4 and 0.8. This underscores the essential role of explicit cross-attention between 2D visual queries and 3D geometric keys for precise multi-object spatial grounding in 3D scene understanding. Crucially, masking the **pose-conditioned dynamic adapter expert** (E_2) causes the most severe overall degradation (a 4.2 drop), with Relative Direction and Route Plan declining by 9.7 and 5.2, respectively. This confirms E_2's essential function in leveraging camera pose to mitigate spatial misalignment from sparse viewpoints, thereby acting as an implicit coordinate transformer to align these disjointed observations. Finally, masking the **gravity-aligned structural expert** (E_3) severely impairs spatial topology and navigation tasks, with Route Plan dropping by 4.0. This demonstrates that our learnable gravity-aligned probes effectively capture the structural foundations and physical anchors necessary for robust spatial grounding. Collectively, these ablation results confirm that each expert fulfills a distinct, specialized role, jointly enhancing the model's comprehensive 3D spatial understanding.

Significance of Multimodal Routing Guidance. Building upon foundational visual and geometric features, our IPAR introduces task instructions and

Table 4: Ablation on expert roles and routing guidance. We analyze each expert ($E_0 - E_3$) via selective masking and evaluate multi-modal routing inputs, where the Base Router relies exclusively on visual and geometric features.

Model Variant	Avg.	Numerical Answer Tasks				Multiple-Choice Answer Tasks			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Ablation on Expert Roles									
w/o E_0	61.8	71.2	46.3	72.9	68.7	57.5	65.6	41.3	70.9
w/o E_1	62.8	70.5	47.3	72.0	69.1	57.3	69.7	43.0	73.5
w/o E_2	59.3	69.8	42.2	72.2	63.0	55.7	60.4	38.8	72.3
w/o E_3	62.4	69.8	47.5	71.7	68.8	58.0	68.6	40.0	74.8
Ablation on Routing Guidance									
Base Router	61.2	70.5	46.0	71.5	67.5	56.5	67.5	39.0	71.1
+ Pose Bias (p)	62.5	70.8	47.2	72.2	68.8	57.8	68.5	42.0	72.7
+ Query Guidance (q)	61.8	70.3	46.5	71.9	68.0	57.2	66.0	41.5	73.0
SpaR3D-MoE (Ours)	63.5	71.3	48.0	72.8	69.6	58.7	70.1	44.0	73.5

Table 5: Ablation on sampling strategy and frame density. We compare ASMS against uniform sampling across varying frame counts on VSI-Bench. ASMS consistently achieves higher performance, with 32 frames delivering the best results.

Sampling Strategy	Avg.	Numerical Answer Tasks				Multiple-Choice Answer Tasks			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Uniform Sampling									
8 Frames	56.6	68.2	42.1	70.0	65.1	53.8	60.1	34.0	59.5
16 Frames	61.2	70.7	47.1	71.8	68.7	57.9	68.1	38.0	67.3
32 Frames	62.4	71.5	47.8	72.6	68.9	58.2	68.4	39.8	72.0
ASMS (Ours)									
8 Frames	57.8	68.5	42.6	70.2	65.5	54.3	62.3	37.6	61.5
16 Frames	61.9	70.8	47.5	71.5	69.3	58.3	69.2	40.1	68.5
32 Frames	63.5	71.3	48.0	72.8	69.6	58.7	70.1	44.0	73.5

camera poses to drive cross-modal interactions for precise expert selection. As shown in the lower section of Tab. 4, integrating these additional modalities significantly optimizes routing decisions. Compared to the base router, task instructions provide semantic guidance with a 0.6 average improvement, while incorporating camera poses yields a larger gain of 1.3. Specifically, pose features directly benefit viewpoint-dependent tasks, increasing Route Plan and Relative Direction by 3.0 and 1.0, respectively. This underscores both modalities as essential condition priors. Their joint synergy culminates in the peak average of 63.5, demonstrating that precise expert dispatching relies on tightly coupling task-aware semantic intent with pose dynamics.

Impact of Sampling Strategy and Frame Density. To evaluate our keyframe extraction mechanism, Tab. 5 compares the ASMS mechanism against uniform sampling across varying frame counts. ASMS consistently outperforms the uniform baseline across all configurations, demonstrating its ability to capture critical spatial structures for 3D understanding. Notably, the 32-frame setup achieves the highest average score of 63.5, highlighted by a 10.6% improvement in the complex Route Plan task. Even with only 16 frames, the model maintains a competitive performance of 61.9, confirming that ASMS effectively filters spatiotemporal redundancy. Although extreme sparsity at 8 frames leads to an overall performance drop, ASMS exhibits greater robustness on complex tasks like Route Plan and Appearance Order. These results indicate that our adaptive sampling successfully retains spatially informative frames, ensuring reliable 3D reasoning even under highly sparse conditions.

5 Conclusion

We introduced SpaR3D-MoE, a novel framework that endows MLLMs with physically-grounded spatial intelligence relying solely on sparse RGB views, obviating the need for 3D-specific data. By sampling sparse keyframes using the proposed ASMS mechanism, the model effectively filters spatiotemporal redundancy while preserving essential scene topology. Building upon this foundation, the introduced HGI-MoE adaptively dispatches multimodal tokens to specialized experts with distinct, tailored cross-modal fusion capacities, fulfilling the diverse requirements of 3D spatial reasoning. Extensive experiments demonstrate that SpaR3D-MoE achieves SOTA performance on the challenging VSI-Bench, ScanQA, and SQA3D benchmarks, confirming the effectiveness and generalizability of our approach across comprehensive spatial understanding tasks. Future work will extend this framework to online video stream spatial reasoning.

6 Acknowledgments

This work was supported by the project "Research of Rapid 3D Digital Acquisition and Reconstruction Technology and Equipment for Cultural Heritage" (No.2024JK4002) and the National Natural Science Foundation of China under Grant Nos. 62572468 and 62402493.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: Adv. Neural Inform. Process. Syst. vol. 35 (2022)
2. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3674–3683 (2018)
3. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19129–19139 (2022)
4. Bai, S., Cai, Y., Chen, R., Chen, K., et al.: Qwen3-vl technical report. ArXiv **abs/2511.21631** (2025)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. ArXiv **abs/2502.13923** (2025)

6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14455–14465 (2024)
7. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 26428–26438. Seattle, WA, USA (2024)
8. Chen, Y., Yang, S., Huang, H., Wang, T., Xu, R., Lyu, R., Lin, D., Pang, J.: Grounded 3d-llm with referent tokens. ArXiv **abs/2405.10370** (2024)
9. Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., He, E., Yin, H., Molchanov, P., Kautz, J., Fan, L., Zhu, Y., Lu, Y., Han, S.: Longvila: Scaling long-context visual language models for long videos (2024), https://arxiv.org/abs/2408.10188
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 24185–24198 (2024)
11. Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y.K., Huang, P., Luo, F., Ruan, C., Sui, Z., Liang, W.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 1280–1297 (2024). https://doi.org/10.18653/V1/2024.ACL-LONG.70
12. Deng, J., He, T., Jiang, L., Wang, T., Dayoub, F., Reid, I.: 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3772–3782 (2025)
13. Fu, R., Liu, J., Chen, X., Nie, Y., Xiong, W.: Scene-llm: Extending language model for 3d visual understanding and reasoning. ArXiv **abs/2403.11401** (2024)
14. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 20482–20494. New Orleans, LA (2023)
15. Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J., et al.: Chat-scene: Bridging 3d scene and large language models with object identifiers. In: Adv. Neural Inform. Process. Syst. vol. 37. Vancouver, BC, Canada (2024)
16. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3d world. In: Int. Conf. Mach. Learn. vol. 235, pp. 20413–20451 (2024)
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., et al.: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
18. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. ArXiv **abs/2401.04088** (2024)
19. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), https://arxiv.org/abs/2408.03326
20. Li, Y., Jiang, S., Hu, B., Wang, L., Zhong, W., Luo, W., Ma, L., Zhang, M.: Uni-moe: Scaling unified multimodal llms with mixture of experts. IEEE Trans. Pattern Anal. Mach. Intell. **47**, 3424–3439 (2024)

21. Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., Jin, P., Huang, J., Zhang, J., Ning, M., Yuan, L.: Moe-llava: Mixture of experts for large vision-language models. ArXiv **abs/2401.15947** (2024)
22. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 26689–26699 (2024)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 34892–34916 (2023)
24. Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., Rao, Y.: Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution (2025), `https://arxiv.org/abs/2409.12961`
25. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. ArXiv **abs/2210.07474** (2022)
26. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning (2025), `https://arxiv.org/abs/2504.01805`
27. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Adv. Neural Inform. Process. Syst. vol. 30, pp. 5099–5108 (2017)
28. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv **abs/2403.05530** (2024)
29. Shen, S., Hou, L., Zhou, Y.Q., Du, N., Longpre, S., Wei, J., Chung, H.W., Zoph, B., Fedus, W., Chen, X., Vu, T., Wu, Y., Chen, W., Webson, A., Li, Y., Zhao, V.Y., Yu, H., Keutzer, K., Darrell, T., Zhou, D.: Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. ArXiv **abs/2305.14705** (2023)
30. Tang, K., Gao, J., Zeng, Y., Duan, H., Sun, Y., Xing, Z., Liu, W., Lyu, K., Chen, K.: Lego-puzzles: How good are mllms at multi-step spatial reasoning? ArXiv **abs/2503.19990** (2025)
31. Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, A., Fergus, R., LeCun, Y., Xie, S.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In: Adv. Neural Inform. Process. Syst. vol. 37 (2024)
32. Ungerleider, L.G., Haxby, J.V.: 'what'and 'where'in the human brain. Current opinion in neurobiology **4**(2), 157–165 (1994)
33. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025). `https://doi.org/10.1109/CVPR52734.2025.00499`
34. Wang, Z., Huang, H., Zhao, Y., Zhang, Z., Zhao, Z.: Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. ArXiv **abs/2308.08769** (2023)
35. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. ArXiv **abs/2505.23747** (2025)
36. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing (2025), `https://arxiv.org/abs/2506.09965`
37. Xue, F., Zheng, Z., Fu, Y., Ni, J., Zheng, Z., Zhou, W., You, Y.: Openmoe: an early effort on open mixture-of-experts language models. In: Int. Conf. Mach. Learn. pp. 55625–55655 (2024)
38. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al.: Qwen3 technical report (2025), `https://arxiv.org/abs/2505.09388`

39. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., et al.: Qwen2 technical report. ArXiv **abs/2407.10671** (2024)
40. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., et al.: Qwen2.5 technical report. ArXiv **abs/2412.15115** (2024)
41. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10632–10643 (2025)
42. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Dong, Y., Liu, S., Zhang, Y., Yang, J., Li, C., Liu, Z.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics. vol. NAACL 2025, pp. 881–916 (2025)
43. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. Trans. Mach. Learn Res. **2025** (2025)
44. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`
45. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. ArXiv **abs/2505.24625** (2025)
46. Zheng, D., Huang, S., Wang, L.: Video-3d llm: Learning position-aware video representation for 3d scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8995–9006 (2025)
47. Zhi, H., Chen, P., Li, J., Ma, S., Sun, X., Xiang, T., Lei, Y., Tan, M., Gan, C.: Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3761–3771 (2025)
48. Zhu, C., Wang, T., Zhang, W., Pang, J., Liu, X.: Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In: Int. Conf. Comput. Vis. pp. 4295–4305 (2025). `https://doi.org/10.1109/ICCV51701.2025.00409`
49. Zhu, F., Wang, H., Xie, Y., Gu, J., Ding, T., Yang, J., Jiang, H.: Struct2d: A perception-guided framework for spatial reasoning in mllms. arXiv preprint arXiv:2506.04220 (2025)
50. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025), `https://arxiv.org/abs/2504.10479`
51. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Int. Conf. Comput. Vis. pp. 2911–2921 (2023)

SpaR3D-MoE: Adaptive 3D Spatial Reasoning from Sparse Views Meets Geometry-Inductive Mixture-of-Experts

Supplementary Material

Haida Feng[1,2], Hao Wei[1,3†], Haolin Wang[1,2], Shiwei Li[1,2], Chade Li[1,2], and Yihong Wu[1,2,3†]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3] YUKUN Intelligent World, Beijing, China
{fenghaida2024@ia, weihao2019@ia, wanghaolin2023@ia, lishiwei2023@ia, lichade2021@ia, yhwu@nlpr.ia}.ac.cn

1 Additional Experimental Details

1.1 Implementation Details

SpaR3D-MoE is built upon the pre-trained Qwen3VL-8B [2] and incorporates VGGT-1B [9] as the visual geometry encoder. During training, both the visual and geometry encoders are frozen, while we optimize the HGI-MoE, multimodal projector, and the MLLM backbone. Specifically, the MLLM backbone is fine-tuned using LoRA [6] with rank $r = 256$ and scaling factor $\alpha = 512$. The network is optimized using AdamW [7] with a weight decay of 0.01. We apply a cosine learning rate scheduler with a 3% warmup ratio, setting the peak learning rate to 5×10^{-5} for the MLLM backbone and 2×10^{-4} for the remaining trainable modules, alongside a global batch size of 128. For the dynamic routing within the HGI-MoE, we configure the Top-K gate to activate 3 experts ($K = 3$) for each multimodal token. To ensure balanced expert routing within the MoE module, we apply an auxiliary load-balancing loss with a coefficient of 0.01. The training inputs consist of non-uniformly sampled sparse keyframes, driving the model to capture and reason over varying spatiotemporal dynamics. For training data, we utilize a mixed dataset consisting of a 288K subset from VSI-590K [11], alongside the training sets of VICA-322K [5], ScanQA [1], and SQA3D [8]. The entire training process runs for one epoch and takes approximately 11 hours on 5 × NVIDIA RTX PRO 6000 Blackwell (96GB) GPUs.

† Corresponding authors

1.2 Evaluation Datasets and Metrics

We utilize the LMMs-Eval framework [13] for standardized evaluation across all benchmarks, extending it with custom evaluation scripts to seamlessly integrate unsupported datasets and ensure unified metric computation. To guarantee deterministic and reproducible results, we employ greedy decoding (temperature $\tau = 0$, beam size 1). For our primary baseline, Qwen3VL, we conduct local evaluations under identical settings, while results for other methods are directly quoted from their original publications to ensure consistency and fairness in comparison. The specific evaluation datasets and metrics are introduced as follows.

VSI-Bench Dataset. VSI-Bench [10] is a comprehensive video-based benchmark designed to evaluate the visual-spatial intelligence of MLLMs. It contains over 5,000 QA pairs derived from 288 real-world indoor egocentric videos sourced from 3D environments like ScanNet [4], ScanNet++ [12], and ARKitScenes [3]. These QA pairs are structured to evaluate 8 fine-grained spatial reasoning tasks, which are categorized into two distinct formats: Multiple-Choice Answer (MCA) tasks and Numerical Answer (NA) tasks. For evaluation metrics, we follow the official protocol, adopting standard Accuracy for MCA tasks and Mean Relative Accuracy (MRA) for NA tasks.

ScanQA Dataset. ScanQA [1] is a foundational 3D visual question answering benchmark designed to assess a model's spatial reasoning and semantic grounding capabilities within complex indoor environments. We conduct our evaluation using the official validation split, which consists of over 4,000 question-answering pairs derived from real-world scenes in the ScanNet dataset [4]. Consistent with the official evaluation protocol, we report Exact Match (EM) accuracy to measure strict answer correctness, and employ standard natural language generation metrics (BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr) to evaluate the fluency and descriptive quality of the generated responses.

SQA3D Dataset. SQA3D [8] is a benchmark dedicated to situated question answering, requiring models to perceive and reason about 3D environments from a specific agent's egocentric position and orientation. We conduct our evaluation on the official test split, comprising over 3,000 question-answering pairs that are classified into six distinct reasoning categories (What, Is, How, Can, Which, and Other). Following the official protocol, we report the Exact Match at 1 (EM@1) accuracy to assess the performance within each individual question category, as well as the overall average score.

2 Additional Ablation Study

2.1 Component Decoupling Analysis

To disentangle the contributions of ASMS and HGI-MoE, we conduct a 2×2 cross-ablation on VSI-Bench by varying the sampling strategy and the inte-

Table 6: Cross-ablation of ASMS and HGI-MoE on VSI-Bench. All variants use 32 input frames.

Sampling	Integration Architecture	Avg. Score	Gain
Uniform	Monolithic Fusion	58.4	–
ASMS	Monolithic Fusion	59.3	+0.9
Uniform	HGI-MoE	62.4	+4.0
ASMS	HGI-MoE	63.5	+5.1

Table 7: Robustness to upstream geometry and pose perturbations on VSI-Bench.

Noise Type	Geo. σ^2	Pose σ^2	Avg.	Δ
Baseline (None)	0.00	0.00	63.5	–
Geo. Noise Only	0.10	0.00	62.6	-0.9
Pose Noise Only	0.00	0.10	61.8	-1.7
Combined Noise	0.10	0.10	62.2	-1.3

gration architecture. Specifically, we compare uniform sampling and ASMS under both monolithic visual-geometric fusion and the proposed HGI-MoE. As shown in Tab. 6, replacing uniform sampling with ASMS improves the sparse keyframe representation by preserving more informative spatiotemporal connectivity, while replacing monolithic fusion with HGI-MoE improves the integration architecture for adaptive multimodal fusion. Combining both components achieves the best performance, suggesting that topology-aware sparse sampling and adaptive heterogeneous integration provide complementary benefits for 3D spatial reasoning.

2.2 Robustness to Geometry and Pose Perturbations

Since SpaR3D-MoE relies on VGGT-predicted geometry features and camera poses for sampling and fusion, we evaluate its sensitivity to upstream prediction noise. Specifically, we inject Gaussian noise with variance $\sigma^2 = 0.1$ into the normalized geometry features and/or pose features extracted by VGGT, and report the results on VSI-Bench in Tab. 7. The model shows a moderate performance drop under both geometry and pose perturbations, indicating that accurate upstream geometry is beneficial for spatial reasoning. Nevertheless, SpaR3D-MoE maintains competitive performance under all perturbation settings, suggesting a certain degree of robustness to imperfect geometry and pose estimates.

Table 8: Ablation on the temporal coefficient ω in ASMS on VSI-Bench. We report average scores for the Numerical and Multiple-Choice Answer categories. All experiments are evaluated via LMMs-Eval [13] using 32 frames.

Method		Avg.	Numerical Answer	Multiple-Choice Answer
Qwen3VL-8B (Baseline)		55.7	**61.9**	49.5
+ ASMS	$\omega = 0.1$	55.1	60.9	49.3
	$\omega = 0.6$ (Default)	**56.2**	61.3	**51.2**
	$\omega = 1.1$	55.6	60.9	50.3

2.3 Hyperparameter Sensitivity in ASMS

To analyze the temporal coefficient ω in ASMS, we extract 32 sparse frames under varying ω values and evaluate them on VSI-Bench using Qwen3VL-8B, comparing against a 32-frame uniform sampling baseline. This hyperparameter controls the balance between geometric diversity and temporal stability. As shown in Tab. 8, the model performance exhibits a clear trade-off, peaking at the intermediate value. Specifically, setting a low $\omega = 0.1$ overly focuses on spatial differences to extract visually diverse frames, but disrupts the temporal progression, reducing accuracy on Multiple-Choice Answer (MCA) tasks (49.3) that require trajectory reasoning. Conversely, setting a high $\omega = 1.1$ makes the temporal penalty dominate, reducing ASMS to a uniform sampling that misses crucial spatial transitions and causes topological disconnection, yielding a average score of 55.6. Notably, both extreme configurations underperform the uniform sampling baseline (55.7), demonstrating that improper temporal constraints degrade spatial reasoning. Ultimately, setting $\omega = 0.6$ achieves the optimal balance. By preserving the spatiotemporal manifold while capturing essential geometric diversity, this setting maximizes the informativeness of the sparse keyframes, resulting in the highest overall average (56.2) and peak MCA accuracy (51.2).

2.4 Efficiency Analysis

We report the system-level efficiency of SpaR3D-MoE on VSI-Bench in Tab. 9. The 3D encoding stage includes VGGT-based feature and pose extraction as well as ASMS keyframe selection, where VGGT extraction takes 7.88s per video and ASMS adds only 0.13s overhead. This scene-level preprocessing makes the initial query take 12.54s, comparable to Spatial-MLLM-4B (12.69s). Within the reasoning stage, the router and MoE introduce only minor overheads of approximately 0.06s and 0.02s, respectively, while SpaR3D-MoE activates about 9B out of 10B total parameters under Top-$K = 3$ routing. Importantly, the preprocessing cost is mainly incurred only when processing a new scene for the first time. For an unchanged scene, the extracted geometry features, camera poses,

Table 9: Latency and memory efficiency on VSI-Bench. The 3D encoding time of SpaR3D-MoE includes VGGT feature extraction and ASMS keyframe selection. Subsequent in-scene question answering can reuse cached scene representations and only requires the reasoning stage.

Model	2D Encoding Time (s)	3D Encoding Time (s)	Reasoning Time (s)	Initial Query Time (s)	Peak Memory (GB)
Spatial-MLLM-4B	0.36	8.93	3.40	12.69	13.10
Qwen3VL-8B	0.72	N/A	2.37	3.09	17.91
SpaR3D-MoE	0.39	8.01	4.14	12.54	27.70

and sparse keyframes can be cached and reused, so subsequent in-scene question answering only requires the reasoning stage and takes approximately 4.14s.

3 Qualitative Analysis

3.1 Qualitative Comparison of Sampling Strategies

We qualitatively compare our ASMS mechanism with the uniform sampling baseline under a sparse 8-frame setting in Fig. 4. For an intuitive comparison, we utilize VGGT [9] to explicitly project the extracted implicit geometric features into a global point cloud, where the corresponding camera poses are marked. As illustrated in Fig. 4(a), uniform sampling extracts frames at fixed temporal intervals, resulting in the omission of critical spatial transitions due to its rigid, content-agnostic nature. Specifically, this topology-agnostic approach fails to capture the doorway transition (indicated by dashed orange boxes) resulting in topological disconnection and a fragmented global point cloud, which compromises global spatial reasoning.

In contrast, Fig. 4(b) demonstrates that ASMS adaptively captures crucial bridge frames (solid green box) by jointly evaluating 3D spatial translations, geometric variance, and temporal dynamics. Preserving these transitional nodes ensures that the extracted sparse frames maintain the underlying topological connectivity of the original scene. Consequently, ASMS is able to provide high-quality foundational scene context, supporting complex downstream spatial tasks, such as relative distance estimation and route planning.

3.2 Qualitative Analysis on 3D Spatial Reasoning

Analysis of Benchmark Scenarios. To further demonstrate the effectiveness of SpaR3D-MoE, we present qualitative comparisons against the Qwen3VL-8B baseline on two distinct spatial reasoning tasks from the VSI-Bench [10]. As illustrated in Fig. 5, we present a Relative Direction task where the agent is required to reason about the spatial relationship between a trash bin and its own location. The baseline Qwen3VL fails to build a consistent 3D representation from sparse 2D views, leading to an incorrect “back-right” prediction. In contrast,

Fig. 4: Qualitative comparison between ASMS and uniform sampling under a sparse 8-frame setting. (a) Uniform sampling misses key spatial transitions (dashed orange boxes), resulting in topological disconnection and a fragmented point cloud. **(b)** Our ASMS adaptively captures crucial bridge frames (solid green box) to preserve the global spatial topology, yielding a coherent scene structure.

our model successfully grounds these 2D observations into a unified 3D scene, enabling it to correctly reason the "front-left" spatial relationship. Furthermore, Fig. 6 showcases a Route Planning task requiring egocentric navigation. The baseline Qwen3VL incorrectly predicts "Turn Right" due to misleading 2D visual information, where the target doorframe appears on the right side of the input image frame. Conversely, SpaR3D-MoE correctly reasons that the doorframe is located behind the agent's current orientation, yielding the accurate "Turn Back" action. This spatial reasoning capability stems from our ASMS, which provides rich scene context while preserving topological connectivity from sparse RGB views, coupled with the HGI-MoE architecture that dynamically activates specialized experts to fuse multimodal geometric and pose features, enabling the agent to construct a unified 3D spatial representation and capture egocentric motion dynamics. By grounding sparse visual inputs into a unified physical space, SpaR3D-MoE overcomes inherent 2D projection ambiguities to perform accurate 3D spatial reasoning.

Generalization to Real-World Scenes. To further validate the zero-shot generalization of SpaR3D-MoE, we conduct qualitative tests on a custom-captured real-world indoor lab scene. As illustrated in Fig. 7, this task requires the model

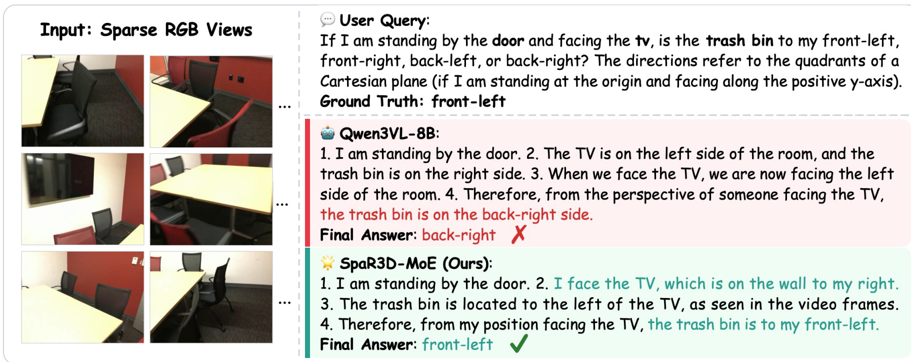

Fig. 5: Qualitative Results on Relative Direction. The baseline Qwen3VL fails to resolve the viewpoint-dependent transformation, leading to spatial hallucinations based on 2D image coordinates. In contrast, SpaR3D-MoE successfully bridges the gap between 2D semantics and 3D geometry, accurately grounding the target's position within the agent's egocentric perspective.

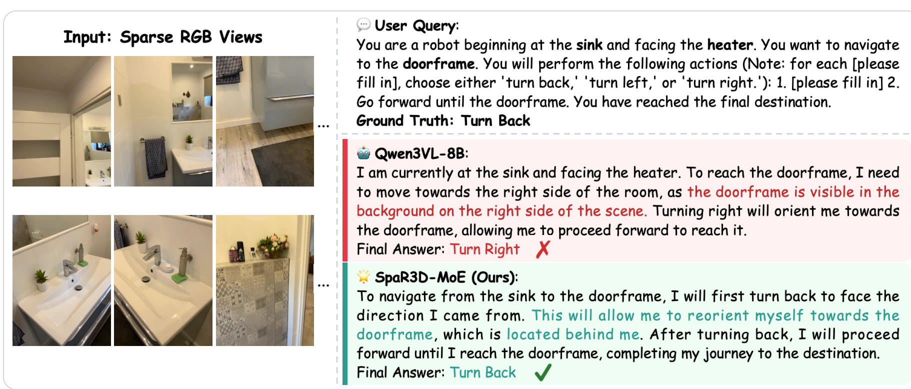

Fig. 6: Qualitative Results on Route Planning. The baseline is biased by the 2D visual information where the target doorframe appears on the right side of the image frame, leading to an incorrect prediction. Conversely, SpaR3D-MoE leverages a consistent egocentric spatial map to correctly reason the agent's egocentric orientation, yielding the accurate "Turn Back" action.

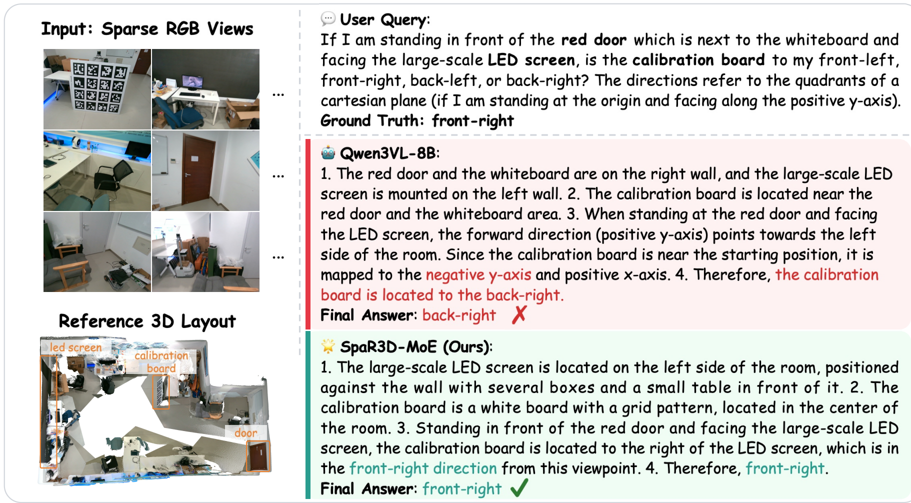

Fig. 7: Qualitative results in an unseen laboratory scene. In an unseen laboratory setting, the baseline incorrectly maps sparse observations to the spatial coordinate system, localizing the target along the negative y-axis. In contrast, SpaR3D-MoE accurately grounds visual cues within a unified physical reference frame, enabling precise reasoning of relative positions from sparse views.

to perform coordinate-based relative direction reasoning within a real-world laboratory scene using only sparse views. The baseline Qwen3VL incorrectly maps sparse 2D visual observations to the spatial coordinate system, leading to a negative y-axis direction and an incorrect "back-right" answer. In contrast, by grounding the red door, LED screen, and calibration board within a unified physical reference frame, our SpaR3D-MoE correctly reasons the "front-right" relative position. These results demonstrate that our method successfully generalizes to real-world scenes using only sparse RGB views, effectively bridging the representational gap between 2D visual semantics and 3D physical geometry to enable robust spatial reasoning.

3.3 Limitations and Failure Analysis

While SpaR3D-MoE demonstrates robust spatial reasoning and generalization capabilities in most scenarios, it still faces challenges with precise instance disambiguation in environments containing repetitive objects. As illustrated in Fig. 8, the model is required to reason about the relative distances to identify the object nearest to the backpack. The thinking step reveals that the model successfully localizes all candidates and correctly associates them with their respective semantic surfaces (e.g., the backpack and refrigerator on the floor, the pillow on the bed). However, the model exhibits a limitation in handling instance ambiguity during distance comparison. Specifically, while the distance to the refrigerator is reasonably estimated ($\sim$1.5m), the distance to the pillow is overestimated ($\sim$3.5m)

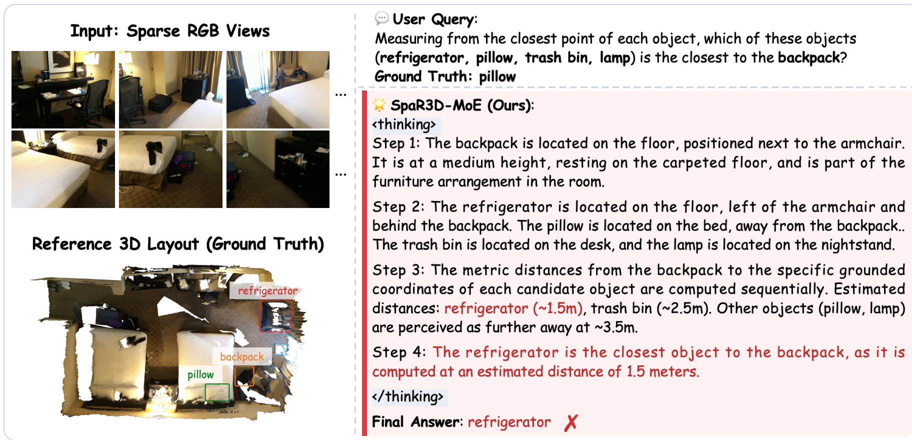

Fig. 8: Qualitative failure case of instance ambiguity. The green box denotes the ground-truth target (pillow), the orange box indicates the reference anchor (backpack), and the red box represents the model's incorrect final prediction (refrigerator). While SpaR3D-MoE correctly identifies semantic categories, it misgrounds the target "pillow" to a distant instance on the opposite bed. Consequently, the distance between the pillow and the backpack is erroneously estimated at ∼3.5m, leading to the incorrect selection of the refrigerator.

because the model incorrectly anchors the semantic token to a distant pillow on the opposite bed. Consequently, this leads to the incorrect final prediction that the refrigerator is the closest object to the backpack. Future improvements will explore integrating Reinforcement Learning (RL) driven by geometric feedback to enhance the model's self-verification capabilities. Cross-validating candidate instances allows the model to calibrate its spatial anchoring, thereby penalizing referential misgrounding and improving instance disambiguation.

References

1. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19129–19139 (2022)
2. Bai, S., Cai, Y., Chen, R., Chen, K., et al.: Qwen3-vl technical report. ArXiv **abs/2511.21631** (2025)
3. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E.: ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
5. Feng, Q.: Visuospatial cognitive assistant. ArXiv **abs/2505.12312** (2025)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017)
8. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. ArXiv **abs/2210.07474** (2022)
9. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025). `https://doi.org/10.1109/CVPR52734.2025.00499`
10. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10632–10643 (2025)
11. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Lu, D., Fergus, R., LeCun, Y., Li, F.F., Xie, S.: Cambrian-s: Towards spatial supersensing in video. ArXiv **abs/2511.04670** (2025)
12. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
13. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Dong, Y., Liu, S., Zhang, Y., Yang, J., Li, C., Liu, Z.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics. vol. NAACL 2025, pp. 881–916 (2025)