Uni3D-MoE: Scalable Multimodal 3D Scene Understanding via Mixture of Experts

Yue Zhang
Zhejiang University

Yingzhao Jian
Zhejiang University

Hehe Fan
Zhejiang University

Yi Yang
Zhejiang University

Roger Zimmermann
National University of Singapore

Abstract

Recent advancements in multimodal large language models (MLLMs) have demonstrated considerable potential for comprehensive 3D scene understanding. However, existing approaches typically utilize only one or a limited subset of 3D modalities, resulting in incomplete representations of 3D scenes and reduced interpretive accuracy. Furthermore, different types of queries inherently depend on distinct modalities, indicating that uniform processing of all modality tokens may fail to effectively capture query-specific context. To address these challenges, we propose Uni3D-MoE, a sparse Mixture-of-Experts (MoE)-based 3D MLLM designed to enable adaptive 3D multimodal fusion. Specifically, Uni3D-MoE integrates a comprehensive set of 3D modalities, including multi-view RGB and depth images, bird's-eye-view (BEV) maps, point clouds, and voxel representations. At its core, our framework employs a learnable routing mechanism within the sparse MoE-based large language model, dynamically selecting appropriate experts at the token level. Each expert specializes in processing multimodal tokens based on learned modality preferences, thus facilitating flexible collaboration tailored to diverse task-specific requirements. Extensive evaluations on standard 3D scene understanding benchmarks and specialized datasets demonstrate the efficacy of Uni3D-MoE.

1 Introduction

3D scene understanding is fundamental for intelligent systems such as robotic navigation [1, 2, 3] and autonomous driving [4, 5, 6, 7]. Recent advances in multimodal large language models (MLLMs) have demonstrated considerable potential for enhancing the interpretation and analysis of complex 3D environments [8, 9, 10, 11, 12, 13]. Usually, existing methods for multimodal 3D scene understanding leverage specific combinations of input modalities. For instance, Chat-3D [14] constructs 3D MLLMs primarily from point clouds. Chat-Scene [15] combines multi-view RGB images and point cloud data. GPT4Scene [16] integrates RGB images of multiple views with bird's eye view (BEV) representations. Video-3D LLM [17] leverages positional video and 3D coordinates to generate spatially-aware representations.

Despite these advancements, leveraging diverse modalities effectively for comprehensive 3D understanding remains challenging, as shown in Fig. 1. Two critical limitations persist: 1) Existing methods usually rely on a limited subset of available modalities, potentially omitting essential information due to occlusions or viewpoint restrictions. For example, relying solely on multi-view RGB images might fail to capture obscured objects, complicating queries such as "How many TVs are in the house?" 2) Different question types exhibit varying dependencies on specific modalities. Queries about object geometry, like "What shape is the wooden desk?", are better addressed with geometric modalities

Preprint. Under review.

such as point clouds, while color-based queries (e.g., "What color is the blanket on the bed?") primarily use visual cues from RGB or depth images. Current dense architectures typically process all modalities uniformly, hindering adaptive alignment with query-specific modality preferences.

In this paper, we propose **Uni3D-MoE**, a scalable and adaptive multimodal 3D scene understanding framework built upon a sparse Mixture-of-Experts (MoE) architecture. Uni3D-MoE comprises three key components: 1) modality-specific encoders that extract features from multiple input modalities, including multi-view RGB and depth images, BEV maps, point clouds, and voxels; 2) modality-alignment adapters that unify these diverse modality-specific features into a common latent representation; and 3) sparse MoE modules integrated within the large language model (LLM), featuring a learnable routing mechanism to dynamically select appropriate experts for each modality token. Our Uni3D-MoE has two merits: 1) by integrating comprehensive 3D modalities, the model achieves more complete and accurate scene representations; 2) the routing mechanism selectively activates relevant expert pathways based on modality tokens, enabling specialized, adaptive processing tailored to each modality.

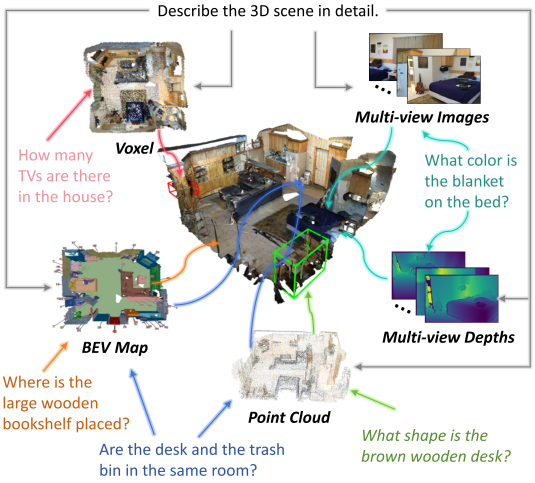

Figure 1: Challenges in 3D scene understanding. (1) Limited modalities may not provide enough scene information. (2) Different question types have varying dependencies on modalities. Existing methods typically treat all modality tokens equally, without adapting to question-specific modality preferences.

Extensive experiments on public 3D scene understanding benchmarks [18, 19, 20] and datasets curated around specific question types demonstrate the effectiveness of Uni3D-MoE. In particular, our model achieves CIDEr gains of 46.9 on the location task and 51.9 on the color task through comprehensive modality details, with further improvements of 6.6 and 10.2 from introducing MoE. The main contributions are summarized as follows:

- **Unified 3D MoE architecture.** We propose Uni3D-MoE, the first unified sparse MoE-based MLLM explicitly designed for 3D scene understanding, supporting a wide range of modalities, including multi-view RGB-D images, BEV maps, point clouds, and voxels.
- **Exploring MoE for Adaptive 3D Modality Fusion.** As the first time, we employ sparse MoE to adaptively fuse 3D modalities. The adaptive routing effectively enhances multimodal fusion tailored to specific queries.
- **Enhanced performance.** Extensive empirical evaluation demonstrates that Uni3D-MoE significantly outperforms existing methods on several 3D scene understanding tasks.

2 Related Work

3D Vision-language Learning. Early research mainly leveraged point clouds [21, 22, 23, 24, 25] or voxels [10, 26, 27] to model objects and scenes, enabling LLMs to perform basic 3D grounding [13, 11, 28, 29] and question answering [30, 31, 32]. Motivated by the spatial ability of self-supervised 2D encoders, subsequent works explore projecting 2D features into 3D space [33, 27], which enhances the comprehension of fine-grained details and complex structures. Recently, 2D videos [10, 34, 16, 17] and BEV [16] have also demonstrated comparable performance to 3D representations, prompting a renewed focus on 2D information in spatial understanding. While these efforts have significantly advanced the field [35, 36, 37, 38, 39], most of them focus on exploiting the strengths of a single modality. The exploration of complementary advantages across multiple modalities remains an open and compelling research topic.

Mixture of Experts. The Mixture-of-Experts (MoE) framework achieves comparable performance to dense models while activating far fewer parameters during inference [40, 41, 42, 43, 44]. MoE is typically categorized into two types based on whether the routing is learned: hard and soft routers. The hard router [45, 46] is suitable for scenarios with clear modality boundaries and predefined expert

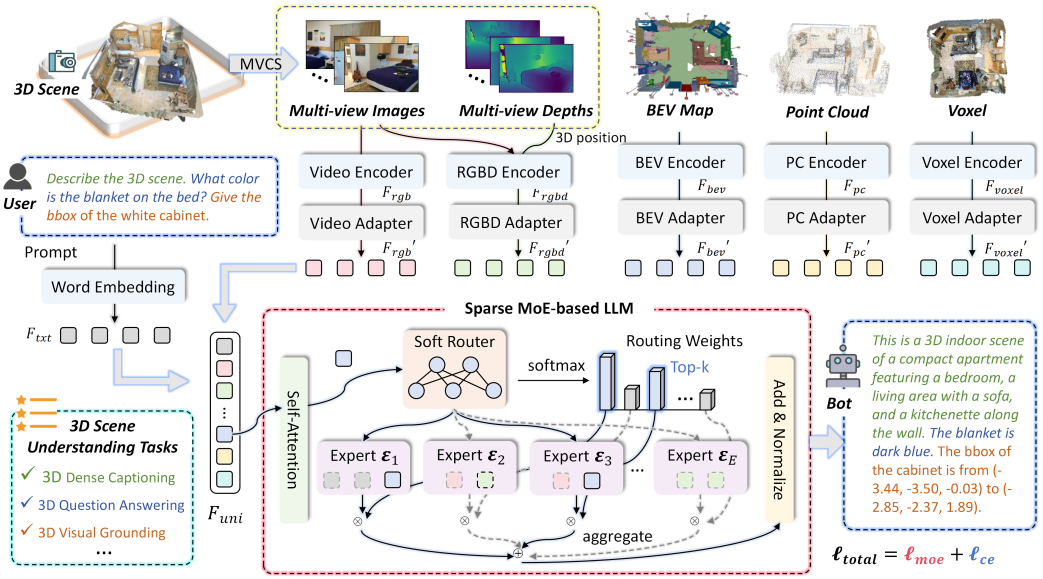

Figure 2: Overview of our method. Uni3D-MoE covers major input modalities of 3D scenes, including RGB/depth images, BEV maps, point clouds, and voxels. To ensure informative spatial coverage, multi-view images are selected using a Maximum Voxel Coverage Sampling (MVCS) algorithm. Each modality is encoded by a modality-specific encoder and aligned via lightweight adapters. The resulting 3D visual tokens, together with text tokens, are then fed into a sparse Mixture-of-Expert (MoE)-based LLM. A learnable soft router dynamically assigns each token to a subset of suitable experts for specialized processing. The model is optimized with a joint objective combining cross-entropy loss l_{ce} and a sparsity-aware expert balancing loss l_{moe}.

assignments. It directly routes different modalities (e.g., images, text) to designated experts [14, 36]. However, hard routing is inflexible, unable to capture token-level semantics or adapt expert selection to tasks [47]. To improve model adaptability, recent dense LLMs have introduced sparse soft routing mechanisms [48, 49], such as LLaMA-MoE [50], MoE-LLaVA [47], and Uni-MoE [51]. These models primarily operate on 1D text and 2D visual inputs. 3D-MoE [52] and MiniGPT-3D [53] initially explore MoE for 3D tasks but lack unified modeling of diverse 3D modalities like voxels, BEV and multi-view images for scene understanding. To this end, we aim to develop a 3D MLLM with soft-routing MoE to achieve unified and adaptive fusion of diverse 3D modalities.

3 Method

Overview. Uni3D-MoE is a unified 3D MLLM framework that leverages sparse MoE for adaptive scene understanding. Fig. 2 illustrates the architecture of Uni3D-MoE, which contains 3D scene feature encoders, feature alignment adapters, and a sparse MoE-enhanced LLM. First, we introduce the 3D scene feature extractor, designed to handle diverse 3D modalities and produce unified feature tokens (Sec. 3.1). Then, we detail the learnable soft router, which selectively activates expert pathways to enable token-level specialization (Sec. 3.2). Finally, we present the training strategy 3.3 and optimization objectives 3.4.

3.1 3D Scene Feature Extractor

Modality Data Preparation. First, we employ the Maximum Voxel Coverage Sampling (MVCS) algorithm to select informative keyframes. Compared with previous approaches [17], our improved MVCS achieves 100× speed-up in computing coverage by using camera poses instead of depth images. Additionally, we enhance frame quality through voxel weighting, depth pruning, and blur image filtering. Further algorithmic details are provided in the Appendix. Then, to provide global spatial context, we render BEV maps with explicit semantic segmentation cues.

Modality-specific Feature Extraction. Uni3D-MoE employs modality-specific encoders to capture more comprehensive representations of 3D scenes. Specifically, multi-view RGB images are encoded using a pre-trained DINOv2 [54] to obtain $\boldsymbol{F}_{rgb}$. For multi-view RGB-D inputs, we first extract 2D patches via CLIP [55], and then integrate corresponding 3D spatial positions derived from depth maps, generating spatially-aware RGB-D features $\boldsymbol{F}_{rgbd}$. We also use DINOv2 [54] to extract BEV features $\boldsymbol{F}_{bev}$. Point clouds, downsampled via Farthest Point Sampling (FPS) [56], are processed through PointNet++ [56], yielding $\boldsymbol{F}_{pc}$. Voxel grids are voxelized and hierarchically encoded by Mask3D's [57] sparse convolutional U-Net to produce $\boldsymbol{F}_{voxel}$.

Modality Feature Alignment. Subsequently, tokens from five modalities are aligned to the text space via respective adapters: $\boldsymbol{F}'_m = \text{Adapter}_m(\boldsymbol{F}_m) \in \mathbb{R}^{N_m \times D_{txt}}$, where $m \in \{\text{rgb}, \text{rgbd}, \text{bev}, \text{pc}, \text{voxel}\}$, N_m is the token count of modality m and D_{txt} is the target embedding dimension. Finally, the text prompt feature $\boldsymbol{F}_{txt}$, combined with modality-aligned features $\boldsymbol{F}'_m$, composes the unified 3D scene representation: $\mathcal{F}_{uni} = \{\boldsymbol{F}_{txt}, \boldsymbol{F}'_{rgb}, \boldsymbol{F}'_{rgbd}, \boldsymbol{F}'_{bev}, \boldsymbol{F}'_{pc}, \boldsymbol{F}'_{voxel}\} \in \mathbb{R}^{N_{uni} \times D_{txt}}$, where $N_{uni} = \sum_m N_m$ denotes the total number of multimodal tokens.

3.2 Soft Routing for Expert Selection

The MoE module employs a learnable soft routing mechanism to achieve intelligent token-to-expert assignment. Given a token $f_i \in \mathcal{F}_{uni}$ and a set of E experts $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_E\}$, a lightweight routing network computes an affinity score $s_i^{(e)}$ between f_i and each expert $\mathcal{E}_e$, where $e \in \{1, 2, \ldots, E\}$. These scores are then normalized into a probability distribution:

$$\pi_i^{(e)} = \frac{\exp(s_i^{(e)})}{\sum_{j=1}^{E} \exp(s_i^{(j)})}, \quad \text{where} \quad s_i^{(e)} = \boldsymbol{w}_e^\top f_i. \tag{1}$$

Here, $\mathbf{w}_e \in \mathbb{R}^D$ denotes the expert-specific routing parameter for expert $\mathcal{E}_e$, and $\pi_i^{(e)}$ represents the routing probability of token f_i to expert $\mathcal{E}_e$. Each token f_i is routed to its top-k experts with the highest probabilities, and the corresponding outputs are aggregated as:

$$\hat{f}_i = \sum_{e \in \mathcal{S}_i} \pi_i^{(e)} \cdot \mathcal{E}_e(f_i), \tag{2}$$

where $\mathcal{S}_i \subseteq \{1, 2, \ldots, E\}$ is the set of top-k selected experts for f_i, and $\mathcal{E}_e(f_i)$ is the output of expert $\mathcal{E}_e$ applied to token f_i.

To balance expert utilization and ensure specialized routing, we incorporate sparsity-aware expert balancing loss, detailed in Sec. 3.4. In this way, Uni3D-MoE achieves adaptive multimodal fusion within each expert, thus accommodating prompt-specific requirements.

3.3 Training Strategy

As shown in Fig. 3, we design a progressive two-stage training strategy for Uni3D-MoE.

Stage I: The goal of this stage is to align 3D visual representations with the textual space, enabling the LLM to capture semantic cues from diverse modalities. During this stage, the modality-specific adapters and LoRA-injected layers within the LLM are jointly trained, while all visual encoders remain frozen except for the spatial-aware RGBD module and the point cloud encoder. The model is trained with complex instructions spanning multiple downstream tasks. We avoid introducing MoE at this stage due to optimization instability when replacing the dense LLM directly. Instead, we refine the model's instruction-following and generation capabilities to prepare for sparse training.

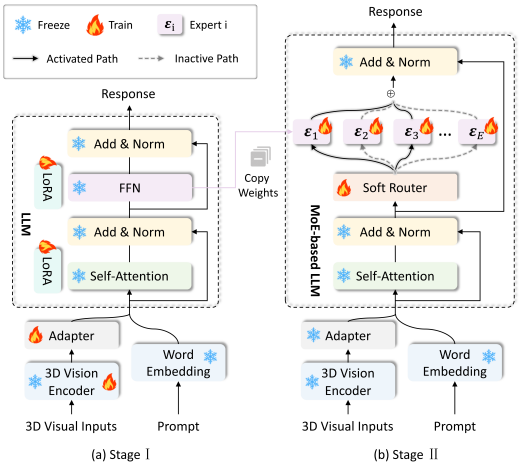

Figure 3: Overview of two-stage training strategy.

Stage II: The objective of this stage is to incorporate the sparse MoE architecture to expand model capacity and enable expert specialization. Inspired by [47], we replicate the feed-forward network (FFN) multiple times to initialize expert modules. At this stage, only the soft router and expert modules are trainable, while all other parameters are kept frozen. To guide the routing process and promote expert diversity, we introduce a sparsity-aware expert balancing loss l_{moe} alongside the standard cross-entropy loss l_{ce}. The training data remains the same as in Stage I, ensuring continuity and stability during the transition to sparse expert routing.

3.4 Training Objective

All tasks are standardized into a unified user-assistant interaction format. During training stage I, the training objective is to minimize the autoregressive cross-entropy loss on the generated text:

$$\mathcal{L}_{\text{ce}} = -\sum_{t=1}^{T} \log P_\theta(\mathcal{Y}_t \mid \mathcal{Y}_{<t}, \mathcal{F}_{uni}), \tag{3}$$

where $\mathcal{Y}_t$ is the t-th target token, $\mathcal{Y}_{<t}$ denotes previously generated tokens, $\mathcal{F}_{uni}$ is the unified multimodal context, and θ represents trainable parameters.

In Stage II, we introduce a sparse MoE mechanism and incorporate sparsity-aware expert balancing loss to encourage expert diversity [49]:

$$\begin{aligned} \mathcal{L}_{\text{moe}} &= E \cdot \sum_{e=1}^{E} \hat{p}^{(e)} \cdot \bar{\pi}^{(e)}, \\ \hat{p}^{(e)} = \frac{1}{N_{uni}} \sum_{i=1}^{N_{uni}} \mathbf{1}\left\{ \arg\max_j \pi_i^{(j)} = e \right\}, &\quad \bar{\pi}^{(e)} = \frac{1}{N_{uni}} \sum_{i=1}^{N_{uni}} \pi_i^{(e)}, \end{aligned} \tag{4}$$

where $\hat{p}^{(e)}$ denotes the fraction of tokens routed to expert e, and $\bar{\pi}^{(e)}$ is the average routing probability to expert $\mathcal{E}_e$. Consistent with the above, E is the number of experts, N_{uni} is the total number of unified tokens, and $\pi_i^{(e)}$ is the routing probability from token i to expert $\mathcal{E}_e$. Here, $\mathbf{1}\{\cdot\}$ denotes the indicator function, which returns 1 if the condition holds and 0 otherwise.

The final training objective is defined as the sum of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{moe}}$, where λ is a balancing coefficient:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda \cdot \mathcal{L}_{\text{moe}}. \tag{5}$$

4 Experiments

4.1 Experiment Settings

Datasets. We construct a unified training corpus by aggregating multiple 3D scene understanding datasets built on ScanNet [58], which contains $1,513$ indoor RGB-D scans with extensive 2D and 3D annotations. The training data covers dense captioning (Scan2Cap [19]), visual question answering (ScanQA[18], SQA3D [20]), and single- and multi-object visual grounding (ScanRefer [59], Multi3DRefer [60]). All data are reformatted into a unified user–assistant interaction format. During evaluation, besides standard 3D scene understanding benchmarks [18, 19, 20], we also use datasets curated around specific question types. Additional details are provided in the Appendix.

Model Details. We initialize our LLM from LLaVA-v1.5-7B [61] and introduce MoE module into layers 8, 12, 16, 20, 24, and 28 at the second training stage. Each MoE layer comprises 8 experts, with the top-2 experts selected for each token at inference time. Following [47, 51], a load-balancing coefficient of $\alpha = 0.01$ is applied to promote expert utilization diversity.

Training Details. We employ a two-stage training strategy: 2 epochs in stage I and 1 epoch in stage II, both with batch size 8. Both stages utilize the AdamW optimizer with a constant learning rate of 2e-5. A warm-up schedule with a warmup ratio of 0.03 followed by cosine decay is applied independently to each stage. The input sequence length is capped at 4096 tokens. To optimize training efficiency and memory usage, we use BF16-based mixed-precision training and leverage DeepSpeed ZeRO-2 offloading.

Table 1: Evaluation results of 3D question answering on ScanQA [18] and SQA3D [20], as well as 3D dense caption on Scan2Cap [19]. EM@1 refers to the top-1 exact match accuracy; BLEU-1, BLEU-4, METEOR, and CIDEr denote text similarity scores between the predicted and ground-truth answer. For Scan2Cap [19], CIDEr is reported at IoU threshold of 0.5. ⋆ indicates that high-resolution settings are not used. We highlight the best performance in **red** and the second-best in **blue**.

Method	ScanQA						SQA3D	Scan2Cap
	EM@1	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	EM@1	CIDEr
Task-specific								
ScanQA [18]	21.1	30.2	10.1	13.1	33.3	64.9	47.2	-
3D-VLP [67]	-	30.5	11.2	13.5	34.5	-	54.9	55.0
3D-VisTA [68]	22.4	-	-	13.9	35.7	-	48.5	66.9
2D LLMs								
InternVL2-8B [69]	-	-	3.3	14.5	34.3	62.5	33.0	-
Qwen2-VL-7B [70]	-	27.8	3.0	11.4	29.3	53.9	40.7	-
LLaVA-Video [71]	-	-	3.1	17.7	44.6	88.7	48.5	-
3D LLMs								
PQ3D [26]	20.0	36.1	-	13.9	-	65.2	47.1	**80.3**
LAMM [72]	-	-	5.8	-	-	42.4	-	-
3D-LLM [73]	20.5	39.3	12.0	14.5	35.7	69.4	-	-
Chat-3D [14]	-	29.1	6.4	11.9	28.5	53.2	-	-
Chat-3D V2 [74]	22.9	38.4	7.3	16.1	40.1	77.1	54.7	-
Chat-Scene [15]	21.6	43.2	14.3	18.0	41.6	87.7	54.6	77.1
LL3DA [75]	-	-	13.5	15.9	37.3	76.8	-	65.2
LLaVA-3D [33]	27.0	-	14.5	**20.7**	**50.1**	**91.7**	**55.6**	79.2
LEO [76]	-	-	11.5	16.2	39.3	80.0	50.0	72.4
Scene-LLM [10]	**27.2**	-	12.0	16.6	40.0	80.0	54.2	37.9
GPT4Scene⋆ [16]	-	**43.4**	**14.6**	17.7	43.6	90.9	-	60.6
Uni3D-MoE (ours)	**30.8**	**43.7**	**17.5**	**19.0**	**47.1**	**97.6**	**57.2**	**85.2**

Evaluation Metrics. Following [33, 15, 62], we adhere to the commonly used metrics to comprehensively evaluate our method across multiple tasks. Specifically, for ScanQA [18], we evaluate the top-1 predicted answers using the exact match accuracy (EM@1), the refined exact match protocol (EM-R@1), F1 score, BLEU-1 [63], BLEU-4 [63], METEOR [64], ROUGE [65], and CIDEr [66]. For SQA3D [20], we use EM@1. For Scan2Cap [19], we combine CIDEr with an IoU threshold of 0.5 between predicted and reference bounding boxes.

4.2 Comparison with State-of-the-art Methods

Comparison Results. Table 1 presents the comparative evaluation results on downstream 3D tasks, including ScanQA [18], SQA3D [20], and Scan2Cap [19] benchmarks. Uni3D-MoE exhibits superior performance on the ScanQA benchmark [19], surpassing existing state-of-the-art methods across multiple metrics. Specifically, it achieves relative improvements of 11.7% on EM@1, 16.8% on BLEU-4, and 6.0% on CIDEr. Furthermore, Uni3D-MoE outperforms LLAVA-3D [33] by 2.7% on EM@1 on SQA3D benchmark [20]. On the Scan2Cap [19] benchmark, Uni3D-MoE achieves an improvement of 5.8% on CIDEr@0.5 compared to the previous advanced method PQ3D [26]. More comprehensive results, including visual grounding evaluations, are provided in the Appendix.

Analysis. The performance benefits from the synergistic effect of heterogeneous modalities, where each modality contributes complementary and modality-specific cues. Furthermore, sparse expert routing further enhances this by adaptively selecting the most relevant modalities for each input, leading to more precise 3D scene understanding.

4.3 Analysis of MoE

Modality Aware Expert Specialization. Figure 4 illustrates the modality token routing behavior across MoE layers from two perspectives: expert-centric and modality-centric. First, the top two rows of Fig. 4 visualize the distribution of modality tokens across 8 experts at each MoE layer. The high proportions of RGB, RGBD, and BEV tokens are attributed to their larger token counts in the input. The expert-wise distributions still capture each expert's modality specialization and selection

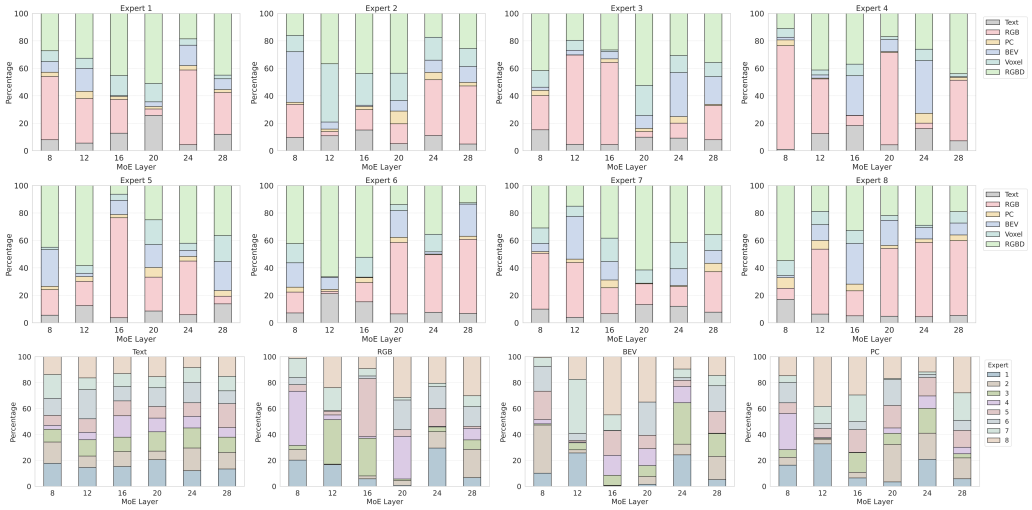

Figure 4: Token-to-expert routing across MoE layers. The first two rows show the modality token distribution across different experts at various MoE layers. The higher proportion of RGB, RGBD, and BEV tokens is attributed to their larger token counts, while expert-wise distributions reveal each expert's modality preferences. The third row presents the expert assignment distribution for each modality, indicating how each modality tends to select different experts throughout the MoE layers.

preferences. For instance, expert $\mathcal{E}_2$ shows a tendency to process voxel and point cloud tokens, which may indicate a specialization in geometric and structural information. Expert $\mathcal{E}_4$ tends to focus more on RGB and BEV inputs, suggesting a possible strength in handling appearance and spatial-view representations. Additionally, expert $\mathcal{E}_6$ at layer 12 and expert $\mathcal{E}_6$ at layer 20 display a noticeable preference for RGBD, potentially reflecting an ability to integrate color and depth cues. Second, the third row of Fig. 4 presents a modality-centric view of expert routing, reflecting which experts usually process tokens of this modality. The results also suggests that multi-view RGB tokens are more often processed by expert $\mathcal{E}_4$, while point cloud tokens tend to be handled by experts $\mathcal{E}_2$, $\mathcal{E}_7$, and $\mathcal{E}_8$. Text tokens appear more evenly distributed across all experts, which may indicate that each expert possesses a basic capacity for processing language information. Overall, the results highlight the effectiveness of our MoE design in promoting modality-aware expert specialization.

Question-type Aware Expert Specialization. To explore expert-specific modality preferences across question types, we select 5 representative categories (color, shape, location, counting, and type), each containing 500 samples from downstream validation/test sets. To mitigate biases caused by differing token counts among modalities, token routing frequencies were normalized within each modality. Fig. 5 provides insights into modality-expert routing preferences across these question categories. Line thickness indicates normalized token routing proportions. Preferred modality-expert routes for each question type are highlighted in color; others are shown in gray. For instance, color-related questions show a preference for the RGB modality tokens routed predominantly to expert $\mathcal{E}_4$, and shape-related questions favor the point cloud modality routed primarily to expert $\mathcal{E}_3$. Additionally, expert $\mathcal{E}_4$ consistently exhibits relatively high activations across multiple question categories, potentially indicating its broader applicability within the MoE structure. These observations demonstrate that our model's routing mechanism exhibits adaptive modality preferences tailored to the specific question.

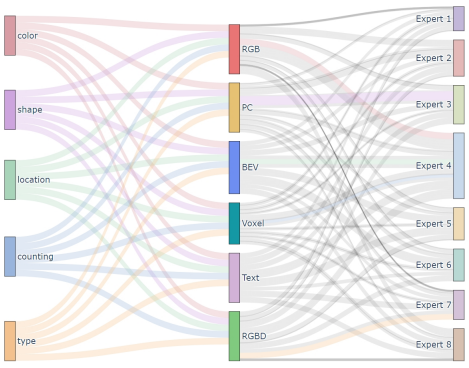

Figure 5: Modality-expert routing preferences across different question types. Line thickness indicates normalized token routing proportions. Preferred modality-expert routes for each query type are highlighted in color; others are shown in gray.

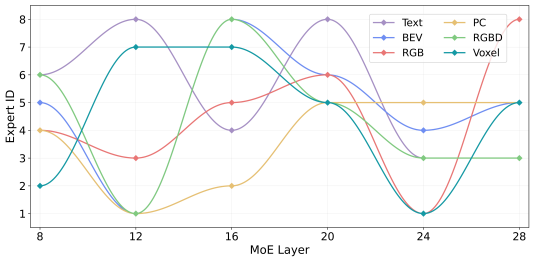

Figure 6: Top-1 activated routing pathways for different modalities, highlighting dynamic and specialized expert activation.

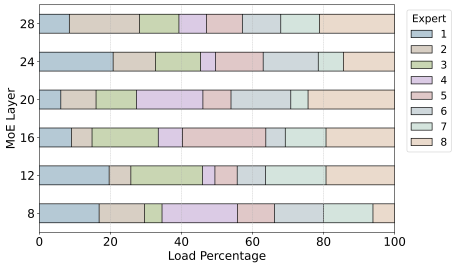

Figure 7: The proportion of tokens assigned to each expert indicates the overall load balance and routing diversity.

Activated Routing Pathways. We further track all tokens on downstream tasks and apply PCA [77] to identify the top-10 most representative routing pathways. Fig. 6 visualizes the top-1 activated expert routing trajectories across MoE layers for each modality. Please refer to the Appendix for more visualization results. We observe that, at layer 20, previously unseen RGB and BEV tokens tend to be routed to expert $\mathcal{E}_6$, whereas PC, voxel, and depth (RGBD) tokens prefer expert $\mathcal{E}_5$. This suggests that Uni3D-MoE may implicitly distinguish visual semantic information from 3D geometric information during high-level representation learning, assigning them accordingly to suitable experts. Additionally, PC and voxel tokens consistently share the same expert across layers 12, 20, and 28, which, to some extent, indicates the model's stable preference for spatial structural features. Overall, these results highlight Uni3D-MoE's modality-aware dynamic routing, enhancing multimodal fusion for scene understanding.

Expert Load Balance. Figure 7 shows the proportion of tokens assigned to each expert across MoE layers. Benefiting from the sparsity-aware expert balancing loss l_{moe}, the overall token distribution remains relatively balanced, which helps improve routing diversity and prevents expert under-utilization.

4.4 Ablation Study

In this section, we conduct ablation studies on Uni3D-MoE. First, we evaluate modality contributions to various question categories. Then, we examine the effectiveness of the MoE module.

Ablation on Modality Contribution. As shown in Table 2, each modality contributes distinctly to different question types. Removing multi-view RGB modality (w/o $\boldsymbol{F}_{rgb}$) primarily impacts performance on the "color" questions, indicating its crucial role in capturing visual color details. Excluding the BEV modality (w/o $\boldsymbol{F}_{bev}$) notably reduces performance on "Location" and "type" tasks, highlighting its importance in spatial understanding. Omitting RGB-D modality (w/o $\boldsymbol{F}_{rgbd}$) mainly decreases performance in the "nature" category, indicating its essential role in capturing detailed characteristics such as object type and shape. Removing point cloud data (w/o $\boldsymbol{F}_{pc}$) considerably affects performance on both "counting" and "nature" categories, demonstrating its strength in capturing object count and structural features. The ablation of voxel modality (w/o $\boldsymbol{F}_{voxel}$) leads to a notable performance drop in "location" and "counting" tasks, underscoring its effectiveness in detailed spatial and quantity understanding. Overall, these results clearly illustrate the complementary and task-specific roles of each modality in Uni3D-MoE, highlighting how their specialized cues collectively facilitate accurate semantic understanding and effective question-specific reasoning.

Ablation on MoE. First, we evaluate the effectiveness of the MoE module across different question categories. Specifically, we select and construct five types of questions (location, type, nature, counting, and color) from downstream tasks. Table 3 demonstrates that incorporating the MoE module consistently improves performance across all these categories, with notable CIDEr gains for "type" and "color". Then, we investigate the effect of varying the number of experts within MoE layers. Considering GPU memory constraints, we set the number of experts per MoE layer to 4, 6, and 8, with each token routed to the top-2 experts. As shown in Table 4, the model with 8 experts per MoE layer outperforms the baseline without MoE by 9.2 on CIDEr. Furthermore, the results show a trend of improved performance with an increasing number of experts, though at the cost of increased training time, suggesting a trade-off between accuracy and computational cost. Additional ablations on the MoE module, including the choice of MoE-equipped layers, are provided in the Appendix.

Table 2: Ablation results of input modalities across five question categories, including location, type, nature, counting, and color, where "nature" mainly covers object shape and type. "w/" denotes experiments using only the specified modality; "w/o" denotes experiments excluding that modality. "w/o 2D info": only point cloud and voxel; "w/o 3D info": only RGB and BEV.

Method	Location		Type		Nature		Counting		Color	
	F1	CIDEr	F1	CIDEr	F1	CIDEr	F1	CIDEr	F1	CIDEr
w/ $\boldsymbol{F}_{rgb}$	25.48	46.42	22.89	40.37	41.64	76.38	45.51	68.03	37.43	62.47
w/ $\boldsymbol{F}_{rgbd}$	20.79	31.50	17.27	31.54	32.34	60.65	37.95	59.66	36.96	57.43
w/ $\boldsymbol{F}_{bev}$	19.92	28.76	18.35	33.56	31.97	56.33	38.94	62.89	33.65	56.92
w/ $\boldsymbol{F}_{pc}$	18.71	26.26	16.95	28.34	33.62	59.31	38.54	62.83	36.73	62.31
w/ $\boldsymbol{F}_{voxel}$	16.00	25.39	18.20	31.57	37.02	69.35	42.12	70.09	31.76	56.85
w/o $\boldsymbol{F}_{rgb}$	32.42	59.89	31.01	55.92	52.25	102.53	53.08	89.53	49.78	86.52
w/o $\boldsymbol{F}_{rgbd}$	34.67	66.34	31.93	55.41	51.25	97.31	51.39	77.83	50.32	87.86
w/o $\boldsymbol{F}_{bev}$	32.43	60.17	30.94	54.85	51.58	98.58	52.65	83.41	49.16	85.75
w/o $\boldsymbol{F}_{pc}$	35.06	69.19	31.73	55.88	50.92	96.86	49.68	75.20	51.59	89.30
w/o $\boldsymbol{F}_{voxel}$	33.16	59.46	32.54	59.21	51.68	99.71	38.58	59.69	51.20	88.68
w/o 2D info	34.33	61.63	30.47	55.31	48.61	91.40	49.82	77.64	48.64	83.40
w/o 3D info	34.33	65.57	31.22	55.31	48.25	89.41	50.33	77.95	51.54	90.80
w/ $\mathcal{F}_{uni}$	37.57	78.59	42.68	80.32	52.89	106.40	50.59	79.52	62.52	111.08

Table 3: Ablation results of MoE module across five question categories, including location, type, nature, counting, and color, where "nature" mainly covers object shape and type.

Method	Location		Type		Nature		Counting		Color	
	F1	CIDEr	F1	CIDEr	F1	CIDEr	F1	CIDEr	F1	CIDEr
w/o MoE	37.5	78.5	42.6	80.3	52.8	106.4	50.5	79.5	62.5	111.0
w/ MoE	39.5$_{\uparrow 2.0}$	85.1$_{\uparrow 6.6}$	47.3$_{\uparrow 4.7}$	93.7$_{\uparrow 13.4}$	53.4$_{\uparrow 0.6}$	107.9$_{\uparrow 1.5}$	53.9$_{\uparrow 3.4}$	86.3$_{\uparrow 6.8}$	67.6$_{\uparrow 5.1}$	121.2$_{\uparrow 10.2}$

Table 4: Ablation results on the number of experts (E) in the MoE module on ScanQA[18]. "w/o MoE" indicates the baseline without MoE. "Time" indicates second-stage MoE training duration.

Method	E	EM@1	EM-R@1	F1	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	Time
w/o MoE	-	27.3	45.1	45.5	41.9	13.9	17.1	43.8	88.4	-
w/ MoE	4	29.5	47.1	47.3	43.2	15.8	18.2	45.7	93.1	$\sim 12h$
w/ MoE	6	29.9	47.8	48.4	43.2	16.4	19.0	45.9	95.5	$\sim 14h$
w/ MoE	8	30.8$_{\uparrow 3.5}$	49.0$_{\uparrow 3.9}$	48.8$_{\uparrow 3.3}$	43.7$_{\uparrow 1.8}$	17.5$_{\uparrow 3.6}$	19.0$_{\uparrow 1.9}$	47.1$_{\uparrow 3.3}$	97.6$_{\uparrow 9.2}$	$\sim 17h$

4.5 Limitations

Despite promising results, Uni3D-MoE exhibits limitations due to token budget constraints and dataset quality. Token limits necessitate modality tokens reduction strategies: multi-view images selected by MVCS may omit critical viewpoints, causing incomplete spatial context; similarly, FPS downsampling reduces point cloud density, compromising fine-grained details. Additionally, performance is affected by blurry images and annotation inaccuracies, introducing noise that impacts precise spatial understanding and object localization tasks.

5 Conclusion

In this paper, we propose Uni3D-MoE, a MoE-based 3D MLLM for comprehensive and adaptive scene understanding. Uni3D-MoE integrates multi-view images, depth, BEV, point clouds, and voxels through modality-specific encoders and a sparse MoE mechanism. By augmenting the LLM with learnable sparse MoE layers, our model adaptively activates specialized experts tailored to each token, enabling dynamic modality fusion aligned with prompt-specific needs. Experimental results show that our method achieves competitive performance on multiple scene understanding tasks.

References

[1] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[2] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in Neural Information Processing Systems*, 37:5285–5307, 2024.

[3] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024.

[4] Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multi-modal data-efficient 3d scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[5] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3347–3355, 2025.

[6] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21634–21643, 2024.

[7] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15077–15087, 2024.

[8] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025.

[9] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In *Forty-first International Conference on Machine Learning*, 2024.

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

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

[12] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024.

[13] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE, 2024.

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

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

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

[17] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. *arXiv preprint arXiv:2412.00493*, 2024.

[18] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.

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

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

[21] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024.

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

[23] Dingning Liu, Xiaoshui Huang, Yuenan Hou, Zhihui Wang, Zhenfei Yin, Yongshun Gong, Peng Gao, and Wanli Ouyang. Uni3d-llm: Unifying point cloud perception, generation and editing with large language models. *arXiv preprint arXiv:2402.03327*, 2024.

[24] Qihang Cao and Huangxun Chen. Objvariantensemble: Advancing point cloud llm evaluation in challenging scenes with subtly distinguished objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1944–1952, 2025.

[25] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26417–26427, 2024.

[26] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2024.

[27] Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, et al. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9247–9255, 2025.

[28] Junjie Fei, Mahmoud Ahmed, Jian Ding, Eslam Mohamed Bakr, and Mohamed Elhoseiny. Kestrel: Point grounding multimodal llm for part-aware 3d vision-language understanding. *arXiv preprint arXiv:2405.18937*, 2024.

[29] Yuan Wang, Ya-Li Li, WU Eastman ZY, and Shengjin Wang. Liba: Language instructed multi-granularity bridge assistant for 3d visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8114–8122, 2025.

[30] Wenxuan Zhu, Bing Li, Cheng Zheng, Jinjie Mai, Jun Chen, Letian Jiang, Abdullah Hamdi, Sara Rojas Martinez, Chia-Wen Lin, Mohamed Elhoseiny, et al. 4d-bench: Benchmarking multi-modal large language models for 4d object understanding. *arXiv preprint arXiv:2503.17827*, 2025.

[31] Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark. *arXiv preprint arXiv:2408.16662*, 2024.

[32] Zechuan Li, Hongshan Yu, Yihao Ding, Yan Li, Yong He, and Naveed Akhtar. Embodied intelligence for 3d understanding: A survey on 3d scene question answering. *arXiv preprint arXiv:2502.00342*, 2025.

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

[34] Haomiao Xiong, Yunzhi Zhuge, Jiawen Zhu, Lu Zhang, and Huchuan Lu. 3ur-llm: An end-to-end multimodal large language model for 3d scene understanding. *arXiv preprint arXiv:2501.07819*, 2025.

[35] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024.

[36] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[37] Xiang Li, Jian Ding, Zhaoyang Chen, and Mohamed Elhoseiny. Uni3dl: A unified model for 3d vision-language understanding. In *European Conference on Computer Vision*, pages 74–92. Springer, 2024.

[38] Taolin Zhang, Sunan He, Tao Dai, Zhi Wang, Bin Chen, and Shu-Tao Xia. Vision-language pre-training with object contrastive learning for 3d scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7296–7304, 2024.

[39] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, et al. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. *arXiv preprint arXiv:2405.10255*, 2024.

[40] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

[41] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024.

[42] Xianzhi Du, Tom Gunter, Xiang Kong, Mark Lee, Zirui Wang, Aonan Zhang, Nan Du, and Ruoming Pang. Revisiting moe and dense speed-accuracy comparisons for llm training. *arXiv preprint arXiv:2405.15052*, 2024.

[43] Xiaoniu Song, Zihang Zhong, Rong Chen, and Haibo Chen. Promoe: Fast moe-based llm serving using proactive caching. *arXiv preprint arXiv:2410.22134*, 2024.

[44] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*, 2024.

[45] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[46] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023.

[47] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947, 2024.

[48] Tongtian Yue, Longteng Guo, Jie Cheng, Xuange Gao, Hua Huang, and Jing Liu. Ada-k routing: Boosting the efficiency of moe-based llms. In *The Thirteenth International Conference on Learning Representations*, 2024.

[49] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[50] Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. Llama-moe: Building mixture-of-experts from llama with continual pre-training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15913–15923, 2024.

[51] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3424–3439, 2025.

[52] Yueen Ma, Yuzheng Zhuang, Jianye Hao, and Irwin King. 3d-moe: A mixture-of-experts multi-modal llm for 3d vision and pose diffusion via rectified flow. *arXiv preprint arXiv:2501.16698*, 2025.

[53] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6617–6626, 2024.

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

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[56] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[57] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023.

[58] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[59] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[60] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023.

[61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

[63] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[64] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[65] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[66] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[67] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023.

[68] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.

[69] Dongchen Lu, Yuyao Sun, Zilu Zhang, Leping Huang, Jianliang Zeng, Mao Shu, and Huo Cao. Internvl-x: Advancing and accelerating internvl series with efficient visual token compression. *arXiv preprint arXiv:2503.21307*, 2025.

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

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

[72] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

[73] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

[74] Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *CoRR*, abs/2312.08168, 2023.

[75] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024.

[76] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2024.

[77] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[78] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16464–16473, 2022.

[79] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11124–11133, 2023.

[80] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022.

[81] Shuo Chen, Tan Yu, and Ping Li. Mvt: Multi-view vision transformer for 3d object recognition. *arXiv preprint arXiv:2110.13083*, 2021.

[82] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021.

[83] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022.

[84] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021.

[85] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1610–1618, 2021.

[86] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021.

[87] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022.

[88] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. 2021.

[89] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022.

[90] Jiaming Chen, Weixin Luo, Xiaolin Wei, Lin Ma, and Wei Zhang. Ham: Hierarchical attention model with high performance for 3d visual grounding. *arXiv preprint arXiv:2210.12513*, 2(3), 2022.

[91] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 3drp-net: 3d relative position-aware network for 3d visual grounding. *arXiv preprint arXiv:2307.13363*, 2023.

[92] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19231–19242, 2023.

[93] Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024.

A Summary

The appendix is organized as follows:

Appendix B:More Results of Uni3D-MoE.
Appendix B.1: Visualization Results.
Appendix B.2: Failure Cases.
Appendix B.3: Detailed quantitative comparison between Uni3D-MoE and other baseline models on various benchmarks, including ScanQA [18], SQA3D [20], ScanRefer [59], Multi3DRefer [60] and Scan2Cap [19].
Appendix C: Additional Results for the MoE Module .
Appendix C.1: Visualization results of MoE, including top-10 activated routing pathways and expert assignment distribution for each modality.
Appendix C.2: Ablation results of MoE, including experiments on MoE layer placement and evaluations across multiple benchmarks.
Appendix D: Limitations and Broader Impacts.
Appendix E: Data Details, including dialogue data format and prompt template.
Appendix F: Model Details, including modality-specific encoders, adapters, and the sparse MoE-based LLM.

B More Results of Uni3D-MoE

B.1 Visualization Results

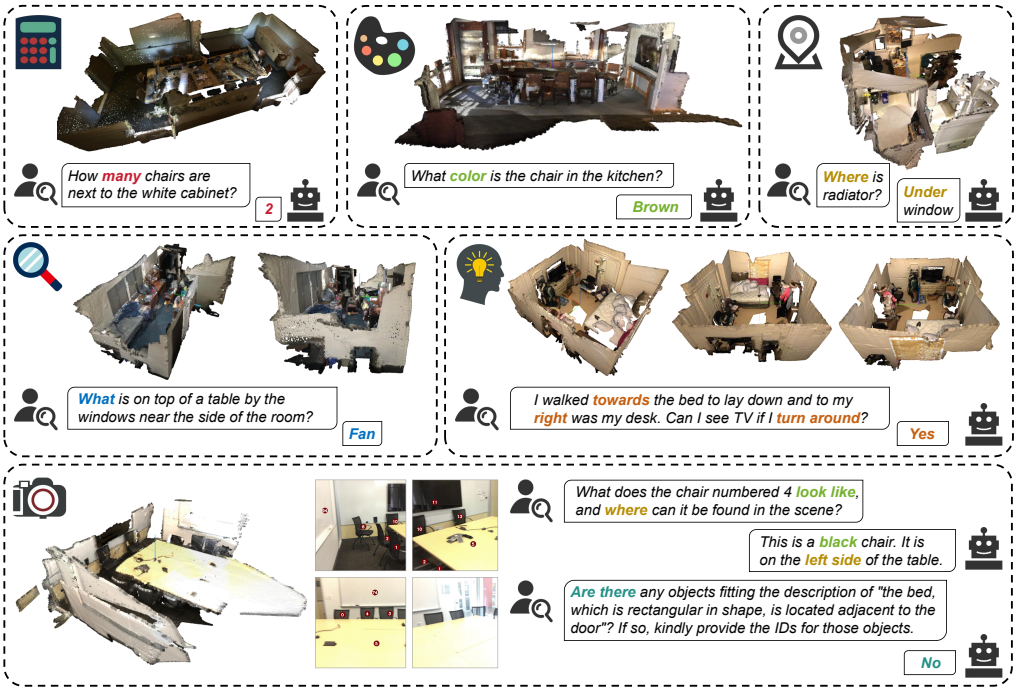

Figure 8: Visualization of Uni3D-MoE performing diverse 3D scene understanding tasks. Examples highlight the model's adaptive multimodal reasoning capabilities across different query types including counting, color recognition, object localization, spatial reasoning, and semantic identification.

As illustrated in Fig. 8, the proposed Uni3D-MoE model adeptly handles various categories of 3D scene understanding tasks, effectively demonstrating its versatility across distinct question types. For instance, it accurately identifies numeric details in counting tasks (e.g., "How many chairs are next to the white cabinet?"), utilizes color recognition to specify attributes (e.g., the "brown" chair in the kitchen), and spatially localizes objects by contextual information (e.g., finding a radiator "under the window"). Moreover, the model is capable of interpreting spatial orientation and viewpoint-dependent questions, successfully answering queries related to turning around to view specific objects. Conversely, it can clearly recognize when queried objects or conditions are absent in the scene, indicating robust negative reasoning capability. These diverse examples underscore

Uni3D-MoE's effectiveness in dynamically leveraging multimodal data representations, showcasing its capability for nuanced and contextually adaptive responses.

B.2 Failure Cases

Fig. 9 illustrates the failure cases of our method. In the first example, given the query "a circular end table, it is next to a teal couch" the model predicts object 13, while the ground truth is object 21. Notably, object 13 also accurately satisfies the description, as it is similarly positioned next to a teal couch and matches the described shape. This indicates that the error arises primarily from inherent annotation ambiguity, rather than from a fundamental shortcoming of the model's visual grounding capability. In the second example, for the query "a black towel, it is hung on the shower curtain rod" the model incorrectly selects object 9 instead of the correct object 13. This failure may be attributed to inconsistent lighting conditions between this scene and others, resulting in color deviations in multi-view RGB images, thus impairing the model's ability to accurately interpret visual cues and distinguish subtle color differences. Additionally, the incorrect prediction might stem from the higher occurrence frequency of object 9 across multiple frames, potentially biasing the model's attention toward it over the less prominently featured yet correct object 13. These cases highlight the importance of addressing both annotation ambiguity and robustness to visual variations in future model improvements.

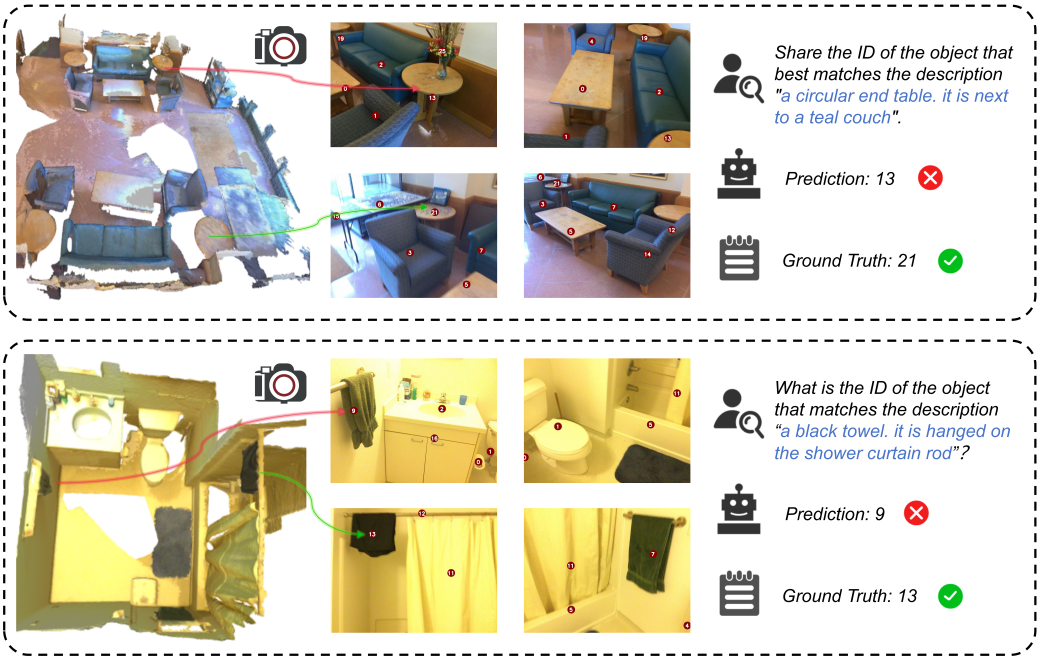

Figure 9: Failure cases of Uni3D-MoE.

B.3 Quantitative Comparison Results

Compared Baselines. We comprehensively evaluate Uni3D-MoE against three categories of state-of-the-art approaches across several 3D benchmarks:

- *Task-specific models* specifically optimized for individual 3D tasks, including ScanQA [18], 3D-VLP [67], 3D-VisTA [68], Scan2Cap [19], 3DJCG [78], Vote2Cap-DETR [79], X-Trans2Cap [80], ScanRefer [59], MVT [81], 3DVG-Trans [82], ViL3DRel [83], and M3DRef-CLIP [60].
- *2D LLMs* adapted from general image-based vision-language models, such as InternVL2-8B [69], Qwen2-VL-7B [70], and LLaVA-Video [71].
- *3D LLMs* that integrate multimodal 3D information into pretrained language models, including PQ3D [26], LAMM [72], Chat-3D [14], Chat-3D V2 [74], 3D-LLM [73], LL3DA [75], LEO [76], Scene-LLM [10], Chat-Scene [15], LLaVA-3D [33], GPT4Scene [16], Ground 3D-LLM [11].

Metric Details. Following previous work [33, 15, 62], we comprehensively evaluate our method using standard metrics across multiple tasks in 3D scene understanding. Specifically:

- For the Scan2Cap [19] task, we assess the quality of generated scene descriptions using widely adopted captioning metrics, including BLEU-4, METEOR, ROUGE, and CIDEr, computed specifically at Intersection-over-Union (IoU) thresholds of 0.25 and 0.5 between predicted and ground-truth bounding boxes.
- For the ScanQA [18] question-answering task, besides captioning metrics, we utilize metrics tailored for answer accuracy and completeness: Exact Match accuracy (EM@1) measures strict correctness of top-1 answers, Relaxed Exact Match (EM-R@1) allows minor acceptable variations, and F1 scores evaluate token-level overlaps.
- For referring expression grounding tasks, ScanRefer [59] and Multi3DRefer [60], we evaluate localization accuracy of predicted bounding boxes against ground-truth annotations. Specifically, we report accuracy (Acc@0.25, Acc@0.5) at IoU thresholds of 0.25 and 0.5 for ScanRefer, and F1 scores (F1@0.25, F1@0.5) at the same IoU thresholds for Multi3DRefer.

In summary, the metrics used can be grouped into three categories: text similarity metrics (BLEU, METEOR, ROUGE, CIDEr) for assessing the quality and fluency of generated descriptions; accuracy metrics (EM@1, EM-R@1, F1) for evaluating exactness and completeness in question-answering tasks; and spatial localization metrics (Acc@IoU, F1@IoU, captioning metrics at IoU thresholds) to quantify the accuracy of bounding-box predictions in scene grounding tasks.

3D Visual Question Answering. Table 5 provides a comprehensive evaluation of various models on the SQA3D benchmark across different 3D question-answering tasks. The tasks are categorized by question types, including "What", "Is", "How", "Can", "Which", and "Others", alongside aggregated metrics of Exact Match accuracy (EM@1) and Relaxed Exact Match accuracy (EM-R@1). Our method, Uni3D-MoE, demonstrates superior performance compared to existing state-of-the-art methods across multiple question types. Specifically, Uni3D-MoE achieves the best results on the "What" (53.1%), "How" (55.8%), "Which" (55.3%), and "Others" (60.2%) question categories, while obtaining second-best performance in "Is" (69.9%). When examining overall accuracy metrics, our model achieves an EM@1 of 57.2% and EM-R@1 of 59.8%, outperforming most baseline methods significantly. These results suggest that our integration of multimodal Mixture-of-Experts (MoE) architecture effectively enhances the model's capacity to process and interpret complex 3D scene queries, particularly in handling open-ended and detailed inquiries.

Table 5: Evaluation results of 3D question answering across different question types on the test set of SQA3D [20]. $\star$ indicates that high-resolution settings are not used. We highlight the best performance in **red** and the second-best in **blue**.

Method	Question Type						Total	
	What	Is	How	Can	Which	Others	EM@1	EM-R@1
Task-specific								
SQA3D [20]	31.6	63.8	46.0	69.5	43.9	45.3	46.6	-
3D-VisTA [68]	34.8	63.3	45.4	69.8	47.2	48.1	48.5	-
ClipBERT [84]	30.2	60.1	38.7	63.3	42.5	42.7	43.3	–
2D LLMs								
InternVL2-8B [69]	30.5	53.8	5.5	47.3	25.8	36.3	33.0	45.3
Qwen2-VL-7B [70]	29.0	59.2	33.4	50.5	44.2	43.2	40.7	46.7
LLaVA-Video-7B [71]	42.7	56.3	47.5	55.3	50.1	47.2	48.5	–
3D LLMs								
LEO [76]	–	–	–	–	–	–	50.0	52.4
Scene-LLM [10]	40.9	69.1	45.0	**70.8**	47.2	52.3	54.2	–
ChatScene [15]	45.4	67.0	52.0	69.5	49.9	55.0	54.6	57.5
LLaVA-3D [33]	–	–	–	–	–	–	55.6	–
GPT4Scene* [16]	**50.7**	**70.9**	48.0	**70.5**	**52.9**	**59.3**	-	**60.7**
Ours	**53.1**	**69.9**	**55.8**	69.5	**55.3**	**60.2**	57.2	**59.8**

3D Visual Grounding. Table 6 summarizes evaluation results for 3D visual grounding tasks on ScanRefer [59] and Multi3DRefer [60], comparing task-specific models and general 3D Large Language Models (3D LLMs). Our method achieves state-of-the-art results, outperforming existing approaches on both benchmarks. Specifically, our model attains the highest accuracy of 62.7% and 57.4% at IoU thresholds of 0.25 and 0.5 respectively on ScanRefer [59], and F1-scores of 65.1% and 60.5% at IoU thresholds of 0.25 and 0.5 on Multi3DRefer [60], demonstrating improvements over the previous best-performing method, GPT4Scene-HDM [16]. These results validate the efficacy of incorporating a Mixture-of-Experts (MoE) architecture into multimodal LLMs, highlighting substantial gains in multimodal grounding capability.

Table 6: Evaluation results of 3D visual grounding on ScanRefer [59] and Multi3DRefer [60]. ⋆ indicates that high-resolution settings are not used. We highlight the best performance in **red** and the second-best in **blue**.

Method	ScanRefer		Multi3DRefer	
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
Task-specific Models				
ScanRefer [59]	37.3	24.3	–	–
MVT [81]	40.8	33.3	–	–
3DVG-Trans [82]	47.6	34.7	–	25.5
ViL3DRel [83]	47.9	37.7	–	–
3DJCG [78]	49.6	37.3	–	26.6
M3DRef-CLIP [60]	51.9	44.7	42.8	38.4
3D LLMs				
3D-LLM [73]	30.3	–	–	–
Ground 3D-LLM [11]	47.9	44.1	45.2	40.6
Chat-Scene [15]	55.5	50.2	57.1	52.4
LLaVA-3D [33]	50.1	42.7	–	–
Ross3D [62]	61.1	54.4	59.6	54.3
GPT4Scene⋆ [16]	40.5	36.7	45.4	42.1
GPT4Scene-HD [16]	50.9	46.4	53.7	50.0
GPT4Scene-HDM [16]	**62.6**	**57.0**	**64.5**	**59.8**
Ours	**62.7**	**57.4**	**65.1**	**60.5**

Table 7 presents a comprehensive comparison between our method and other state-of-the-art approaches on the ScanRefer [59]. Performance is assessed separately across "Unique", "Multiple", and combined "Overall" subsets. The "Unique" subset involves unambiguous samples, each with only a single instance per object category, whereas the "Multiple" subset includes ambiguous samples containing multiple instances from the same category. Metrics used are accuracy measured at IoU thresholds of 0.25 and 0.5. The proposed method achieves superior performance, especially in handling ambiguous cases within the "Multiple" subset, obtaining promising accuracy scores at 56.7% (Acc@0.25) and 51.5% (Acc@0.5). It also demonstrates outstanding overall capabilities, achieving state-of-the-art results on the "Overall" subset with accuracies of 62.7% and 57.4%, closely surpassing the previously best-performing model GPT4Scene-HDM. In the "Unique" subset, our method achieves competitive results (89.6% at Acc@0.25 and 83.5% at Acc@0.5), second only slightly to GPT4Scene-HDM, reflecting strong capability in handling clear, well-defined visual grounding scenarios. These results highlight substantial effectiveness of our method in visual grounding, notably its capability in resolving ambiguity inherent in challenging multi-instance scenes, thus underscoring the advantages brought by integrating Mixture-of-Experts architecture within multimodal large language models.

Table 8 illustrates the comprehensive evaluation results for 3D visual grounding performance on the Multi3DRef [60] across five distinct scenarios: Zero Target without Distractors (ZT w/o D), Zero Target with Distractors (ZT w/ D), Single Target without Distractors (ST w/o D), Single Target with Distractors (ST w/ D), and Multi-Target (MT). Performance is assessed through F1 scores at IoU thresholds of 0.25 and 0.5, emphasizing precision in object localization under varying complexity and distractor presence conditions. The proposed approach demonstrates competitive performance across multiple scenarios, achieving notable results especially in scenarios involving distractors. For instance, our method achieves the highest F1@0.25 (60.0) and F1@0.5 (55.1) scores in the challenging Single Target with Distractors (ST w/ D) scenario, surpassing previous strong models such as GPT4Scene-HDM [16]. Similarly, in the comprehensive evaluation across all scenarios (denoted "ALL"), our method attains leading performance (F1@0.25: 65.1, F1@0.5: 60.5), indicating its broad effectiveness in diverse grounding contexts. Task-specific methods, such as M3DRef-CLIP [60] and 3DICG (Grounding) [78], exhibit strong performance in simpler settings (e.g., ZT w/o D and ST w/o D), though their results show noticeable declines when encountering scenarios with distractors or multiple targets. In contrast, the proposed approach demonstrates enhanced robustness and flexibility in addressing increased task complexity. This observation suggests that explicitly modeling multi-modal complexity and integrating Mixture-of-Experts (MoE) module within LLM frameworks may positively influence grounding performance.

3D Dense Captioning. Table 9 presents the evaluation results of 3D dense captioning on the Scan2Cap [19] benchmark, comparing our model against several state-of-the-art methods. Performance is measured using widely adopted captioning metrics—BLEU-4, METEOR, ROUGE, and CIDEr—at IoU thresholds of 0.25 and 0.5, indicating the quality and spatial accuracy of generated captions.

Our method achieves superior performance compared to other advanced approaches, including both task-specific models and recent 3D LLMs. Specifically, at IoU=0.25, our model attains the highest BLEU-4 (44.4), METEOR (29.9), and CIDEr (89.9) scores, indicating strong fluency, semantic alignment, and relevance of the captions. At a stricter threshold of IoU=0.5, our model also demonstrates leading performance with top results in BLEU-4 (41.1) and CIDEr (85.2), highlighting the method's robustness in precise localization conditions. These findings underscore the advantage of integrating the Mixture-of-Experts architecture into multimodal language models, improving caption generation in complex 3D scenarios.

Table 7: Full Evaluation of 3D visual grounding on ScanRefer [59]. The "Unique" subset contains samples in which the described object corresponds to exactly one unique instance within a given object category, whereas the "Multiple" subset includes ambiguous cases with multiple instances belonging to the same object category. The "Overall" category aggregates performance across both unique and multiple-instance subsets. Accuracy is measured using IoU thresholds of 0.25 and 0.5 between predicted and ground-truth bounding boxes. ⋆ indicates that high-resolution settings are not used. We highlight the best performance in **red** and the second-best in **blue**.

Method	Unique		Multiple		Overall	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Task-specific Models						
ScanRefer [59]	76.3	53.5	32.7	21.1	41.2	27.4
TGNN [85]	68.6	56.8	29.8	23.2	37.4	29.7
X-Trans2Cap [80]	73.2	50.8	37.6	25.2	44.5	30.1
InstanceRefer [86]	75.7	64.7	29.4	23.0	38.4	31.1
3DVG-Trans [82]	81.9	60.6	39.3	28.4	47.6	34.7
MVT [81]	77.7	66.4	31.9	25.3	40.8	33.3
3D-SPS [87]	84.1	66.7	40.3	29.8	48.8	37.0
ViL3DRel [83]	81.6	68.6	40.3	30.7	47.9	37.7
3DJCG [78]	83.5	64.3	41.4	30.8	49.6	37.3
D3Net [88]	–	72.0	–	30.1	–	37.9
BUTD-DETR [89]	84.2	66.3	46.6	35.1	52.2	39.8
HAM [90]	79.2	67.9	41.5	34.0	48.8	40.6
3DRP-Net [91]	83.1	67.7	42.1	32.0	50.1	38.9
3D-VLP [67]	84.2	64.6	43.5	33.4	51.4	39.5
EDA [92]	85.8	68.6	49.1	37.6	54.6	42.3
M3DRef-CLIP [60]	85.3	77.2	43.8	36.8	51.9	44.7
3D-VisTA [68]	81.6	75.1	43.7	39.1	50.6	45.8
ConcreteNet [93]	86.4	82.1	42.4	38.4	50.6	46.5
3D LLMs						
Chat-Scene [15]	89.6	82.5	47.8	42.9	55.5	50.2
Video-3D-LLM [17]	88.0	78.3	50.9	45.3	58.1	51.7
Ross3D [62]	87.2	77.4	54.8	48.9	61.1	54.4
GPT4Scene⋆ [16]	65.5	61.2	34.8	31.1	40.5	36.7
GPT4Scene-HD [16]	77.5	71.9	44.9	40.6	50.9	46.4
GPT4Scene-HDM [16]	**90.3**	**83.7**	**56.4**	**50.9**	**62.6**	**57.0**
Ours	**89.6**	**83.5**	**56.7**	**51.5**	**62.7**	**57.4**

C Additional Results for the MoE Module

C.1 Visualization Results of MoE

Top-10 Activated Routing Pathways. As illustrated in Fig. 10, we visualize the top-10 activated routing pathways across different modalities (Text, RGB, BEV, RGBD, PC, and Voxel) through multiple Mixture-of-Experts (MoE) layers. Each modality is represented in a distinct color, with the most prominent paths (Top-1 and Top-2) highlighted, while the other pathways are depicted in gray to emphasize relative activation strengths. The visualization reveals dynamic and specialized expert activations that vary across layers, highlighting the model's adaptive routing mechanism. For instance, point cloud (PC) modality prominently engages expert $\mathcal{E}_5$ at deeper layers (layers 20, 24, and 28), whereas RGB modality dynamically shifts its primary expert from expert $\mathcal{E}_4$ at layer 8 to expert $\mathcal{E}_1$ at layer 24. This behavior underscores the importance of employing MoE architectures to dynamically allocate modality-specific information to the most suitable experts at different representation depths, thereby enhancing overall model performance.

Expert Assignment Distribution for Each Modality. Fig. 11 illustrates the varying patterns of expert assignment for each modality (Text, BEV, RGB, RGBD, PC, and Voxel) across different MoE layers (layers 8, 12, 16, 20, 24, and 28) within the Uni3D-MoE model. Observations suggest potential modality-specific preferences and evolving trends in expert usage as model depth increases. For example, the RGB modality distinctly varies its expert selection patterns across layers: prominently activating expert $\mathcal{E}_4$ at layer 8, experts $\mathcal{E}_3$ and $\mathcal{E}_5$ at layers 12 and 16, and shifting focus towards experts $\mathcal{E}_6$ and $\mathcal{E}_8$ in deeper layers (layers 20, 24, and 28). This progression may reflect changing requirements in visual feature extraction as information abstraction deepens.

Table 8: Full evaluation results of 3D visual grounding on Multi3DRef [60]. Performance is assessed across five scenarios: Zero Target without Distractors (ZT w/o D), where no object matches the referring expression and no distractors exist; Zero Target with Distractors (ZT w/ D), where no object matches but distractors are present; Single Target without Distractors (ST w/o D), referring to a single, uniquely identifiable target object without distractors; Single Target with Distractors (ST w/ D), a single target object with multiple distractors present; and Multi-Target (MT), where multiple objects match the referring expression simultaneously. Metrics reported include the F1 score (F1) at IoU thresholds of 0.25 and 0.5 (F1@0.25, F1@0.5), reflecting localization precision and recall. "ALL" aggregates results across all five scenarios. ⋆ indicates that high-resolution settings are not used. We highlight the best performance in **red** and the second-best in **blue**.

Method	ZT w/o D	ZT w/ D	ST w/o D		ST w/ D		MT		ALL	
	F1	F1	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5
Task-Specific Model										
3DVG-Trans [82]	87.1	45.8	–	27.5	–	16.7	–	26.5	–	25.5
D3Net (Grounding) [88]	81.6	32.5	–	38.6	–	23.3	–	35.0	–	32.2
3DJCG (Grounding) [78]	94.1	66.9	–	26.0	–	16.7	–	26.2	–	26.6
M3DRef-CLIP [60]	81.8	39.4	53.5	47.8	34.6	30.6	43.6	37.9	42.8	38.4
3D LLMs										
Chat-Scene [39]	90.3	62.6	82.9	75.9	49.1	44.5	45.7	41.1	57.1	52.4
GPT4Scene⋆ [16]	85.2	61.4	60.1	55.1	37.7	34.4	39.4	36.3	45.4	42.1
GPT4Scene-HD [16]	93.6	81.8	72.5	66.2	46.6	42.9	41.8	38.9	53.7	50.0
GPT4Scene-HDM [16]	**97.4**	**84.4**	**85.0**	**77.7**	**59.9**	**55.1**	**48.6**	**44.6**	**64.5**	**59.8**
Ours	**96.8**	**84.7**	**84.9**	**77.3**	**60.0**	**55.1**	**51.4**	**47.7**	**65.1**	**60.5**

Table 9: Evaluation results of 3D dense captioning on Scan2Cap [19]. BLEU-4, METEOR, ROUGE and CIDEr denote text similarity scores between the predicted answer and the ground-truth answer. Metrics are computed under IoU thresholds of 0.25 and 0.5 between the predicted and reference bounding boxes. ⋆ indicates that high-resolution settings are not used. We highlight the best performance in **red** and the second-best in **blue**.

Method	**Scan2Cap** (IoU@0.25)				**Scan2Cap** (IoU@0.5)			
	BLEU-4	METEOR	ROUGE	CIDEr	BLEU-4	METEOR	ROUGE	CIDEr
Task-specific Models								
Scan2Cap [19]	34.2	26.3	55.3	56.8	22.4	21.4	43.5	35.2
3DJCG [78]	40.2	27.7	59.2	64.7	31.5	24.3	51.8	47.7
3D-VLP[67]	41.0	28.1	59.7	70.7	32.3	24.8	51.5	54.9
3D-VisTA [68]	36.5	28.4	57.6	71.0	34.0	26.8	54.3	61.6
Vote2Cap-DETR [79]	39.3	28.3	59.3	71.5	34.5	26.2	54.4	61.8
X-Trans2Cap [80]	35.7	26.6	54.7	61.8	25.1	22.5	45.3	43.9
3D LLMs								
LEO [76]	-	-	-	-	38.2	27.9	58.1	72.4
LL3DA [75]	41.4	27.8	59.5	74.2	36.8	26.0	55.1	65.2
Chat-Scene [15]	38.2	29.0	60.6	81.9	36.3	28.0	58.1	77.1
LLaVA-3D [33]	-	-	-	-	41.1	30.2	63.4	79.2
Ross3D [62]	-	-	-	-	43.4	30.3	66.9	81.3
GPT4Scene⋆ [16]	36.3	26.5	57.6	63.8	34.2	25.6	55.2	60.6
GPT4Scene-HD [16]	40.4	28.3	60.2	79.1	37.9	27.3	57.7	74.4
GPT4Scene-HDM [16]	**43.1**	**29.3**	**61.9**	**91.7**	**40.6**	**28.2**	**59.3**	**86.3**
Ours	**44.4**	**29.9**	**60.7**	**89.9**	**41.1**	**29.9**	**59.3**	**85.2**

Similarly, the BEV modality noticeably engages experts $\mathcal{E}_2$, $\mathcal{E}_7$ and $\mathcal{E}_8$ frequently at intermediate layers but seems to diversify at deeper layers, possibly due to increasing complexity in spatial reasoning tasks. The Voxel modality frequently utilizes experts $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_7$, which might indicate specific geometric feature processing demands. The PC modality prominently selects experts $\mathcal{E}_1$ and $\mathcal{E}_8$ in the initial layers (layers 8, 12, and 16), while in deeper layers (20, 24, and 28) it shifts predominantly towards experts $\mathcal{E}_2$ and $\mathcal{E}_8$. This pattern might indicate evolving requirements in geometric or structural feature extraction as the model processes point cloud data at different abstraction levels. The Text and RGBD modalities exhibit relatively uniform expert distribution patterns across layers, suggesting stable and balanced processing demands possibly related to semantic and visual-depth integration tasks. These patterns collectively highlight Uni3D-MoE's capability to potentially adapt routing strategies for different modalities, thus possibly enhancing multimodal representation effectiveness and task-specific performance.

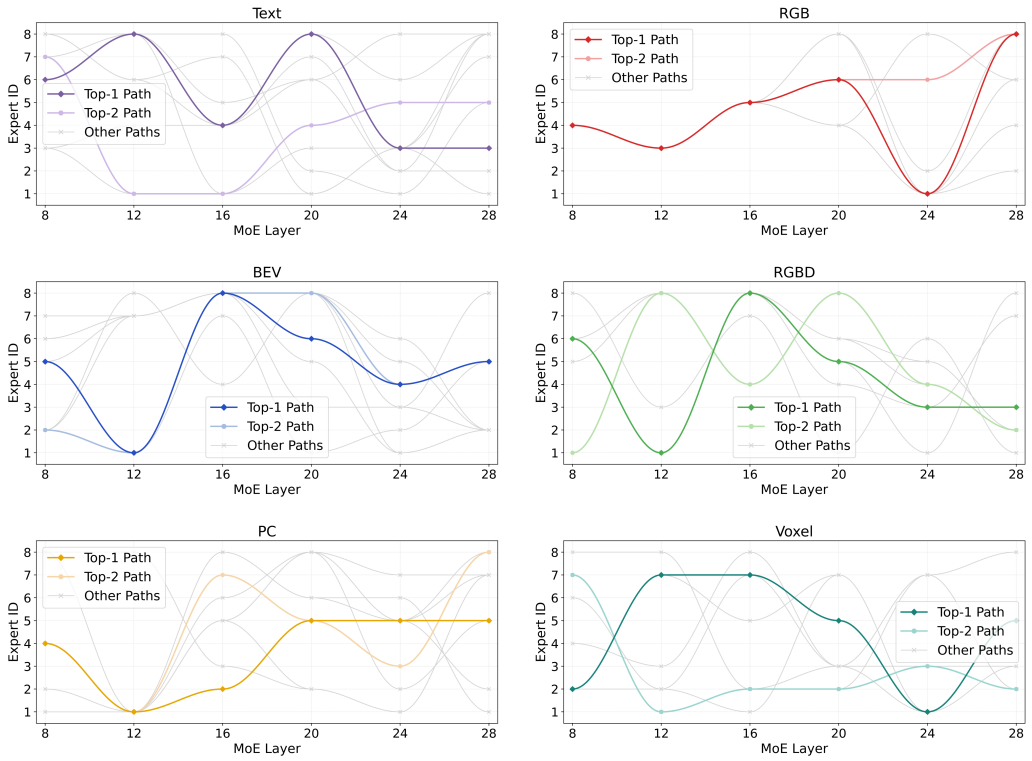

Figure 10: Top-10 activated routing pathways for different modalities, highlighting dynamic and specialized expert activation. Colored curves illustrate the top-1 and top-2 routing paths for each modality, while gray curves represent the remaining eight pathways.

Table 10: Ablation results of MoE layers on ScanQA[18].

Method	MoE layers	EM@1	EM-R@1	F1	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
w/o MoE	-	27.3	45.1	45.5	41.9	13.9	17.1	43.8	88.4
w/ MoE	[0,2,4,6,8,10]	27.7	45.3	46.2	41.5	14.1	17.6	44.0	90.9
w/ MoE	[0,4,8,12,16,20]	28.7	47.1	47.8	42.2	15.7	17.9	46.0	94.3
w/ MoE	[8,12,16,20,24,28]	30.8	49.0	48.8	43.7	17.5	19.0	47.1	97.6

C.2 Ablation Results of MoE

Ablation on MoE Layer Placement. Table 10 presents the ablation results for integrating the Mixture-of-Experts (MoE) module at different layers in the ScanQA [18] benchmark. We observe a clear performance improvement across all evaluation metrics when employing the MoE module at deeper layers. Specifically, incorporating MoE layers at depths [8,12,16,20,24,28] achieves the best results, significantly enhancing EM@1 accuracy from 27.3% (without MoE) to 30.8%. Similar improvements are evident across other metrics, such as CIDEr, which rises notably from 88.4 to 97.6. These results highlight that deeper integration of the MoE mechanism facilitates richer feature extraction, thereby enhancing the model's ability to accurately answer complex questions.

MoE Ablation Across Benchmarks. Tables 11-13 present additional ablation results of the Mixture-of-Experts (MoE) module on different 3D scene understanding tasks. Specifically, on the SQA3D [20] benchmark, integrating MoE yields clear improvements, achieving better exact-match accuracy (EM@1: 57.2 vs. 54.6) and significantly higher text similarity scores (e.g., CIDEr: 147.8 vs. 136.0). For visual grounding tasks on ScanRefer [59] and Multi3DRefer [60], the MoE-equipped GPT4Scene backbone consistently shows performance gains (ScanRefer Acc@0.5: 57.4 vs. 57.0; Multi3DRefer F1@0.5: 60.5 vs. 59.8). In the 3D dense captioning scenario (Scan2Cap [19]), the MoE integration brings mixed results, slightly improving BLEU-4 and METEOR at IoU@0.25 but slightly decreasing CIDEr scores. Overall, the introduction of MoE demonstrates consistent, albeit varied, improvements across multiple 3D scene understanding tasks.

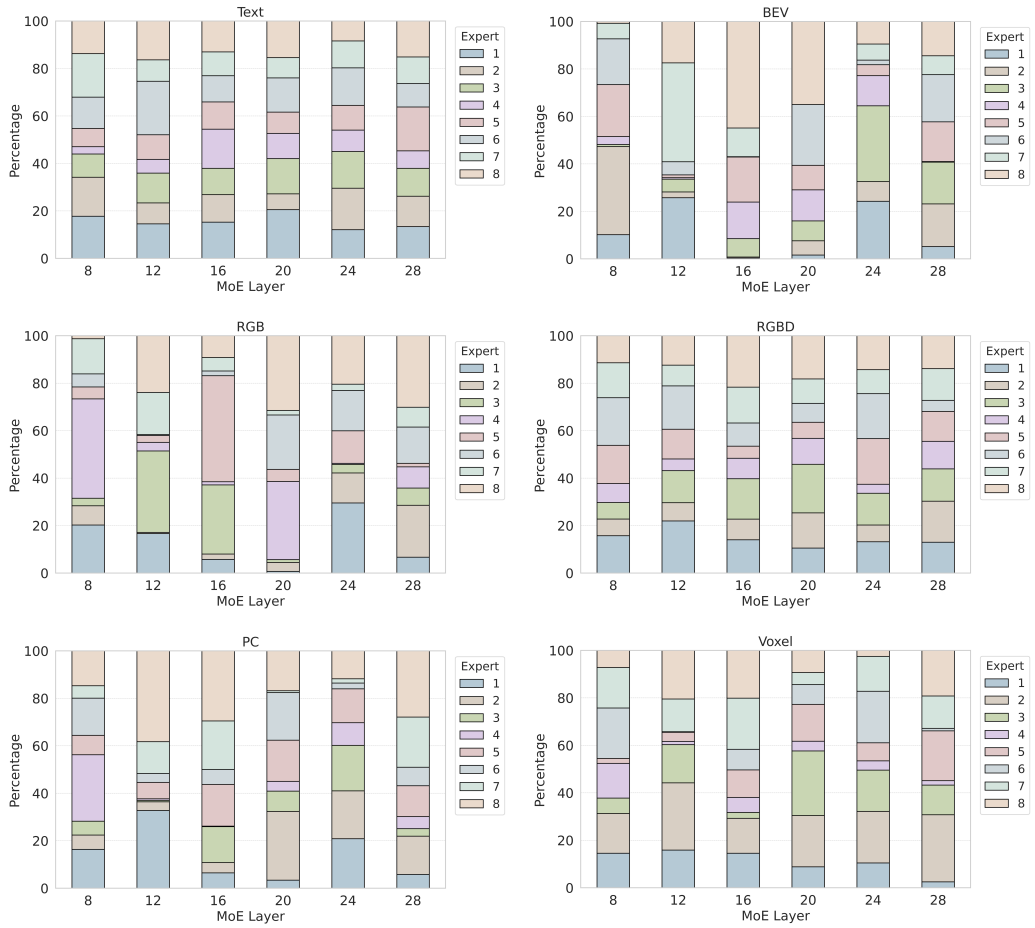

Figure 11: Expert assignment distribution for each modality across MoE layers in Uni3D-MoE, highlighting modality-specific expert selection dynamics.

Table 11: Ablation results of the MoE module on SQA3D test set [20] for 3D question answering. EM@1 refers to the top-1 exact match accuracy; BLEU-1, BLEU-4, METEOR, and CIDEr denote text similarity scores between the predicted and ground-truth answer.

Method	EM@1	EM-R@1	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
w/o MoE	54.6	55.7	50.4	39.6	35.3	52.8	136.0
w/ MoE	57.2	59.8	54.9	43.5	38.3	57.9	147.8

Table 12: Ablation results of the MoE module on ScanRefer [59] Multi3DRefer [60] for 3D visual grounding.

Method	**ScanRefer**		**Multi3DRefer**	
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
w/o MoE	62.6	57.0	64.5	59.8
w/ MoE	62.7	57.4	65.1	60.5

D Limitations and Broader Impacts

Limitations. Despite achieving promising results across various tasks, Uni3D-MoE still exhibits several limitations. First, the token budget constraints of large language models necessitate strict control over modality-

Table 13: Ablation results of the MoE module on Scan2Cap [19] for 3D dense captioning. BLEU-4, METEOR, and CIDEr denote text similarity scores between the predicted answer and the ground-truth answer. Metrics are computed under IoU thresholds of 0.25 and 0.5 between the predicted and reference bounding boxes. ⋆ indicates that high-resolution settings are not used.

Method	**Scan2Cap** (IoU@0.25)				**Scan2Cap** (IoU@0.5)			
	BLEU-4	METEOR	ROUGE	CIDEr	BLEU-4	METEOR	ROUGE	CIDEr
w/o MoE	43.1	29.3	61.9	91.7	40.6	28.2	59.3	86.3
w/ MoE	44.4	29.9	60.7	89.9	41.1	29.9	59.3	85.2

specific inputs. To this end, multi-view images are selected using the Maximum Voxel Coverage Sampling (MVCS) algorithm. While effective, this method may overlook critical viewpoints, leading to an incomplete spatial context. Similarly, point clouds are downsampled using Farthest Point Sampling (FPS), reducing point density and limiting the representation of fine-grained object details—particularly for small-scale structures. These input reductions might degrade model performance, especially when other modalities fail to provide sufficient complementary information. Second, the model's effectiveness is partially constrained by the quality of the training dataset. Blurry multi-view images and annotation inaccuracies introduce noise and ambiguity, which can hinder performance in tasks that demand precise spatial understanding and accurate object localization.

Broader Impacts. This paper aims to enhance the 3D perception capabilities of VLM. The proposed Uni3D-MoE has potential applications in human-computer interaction and autonomous robotics. It can help embodied agents better understand the environment and perform complex tasks. While there are potential concerns about misuse, such as applications in military robotics, we believe the benefits of our approach significantly outweigh the minimal risks.

E Data Details

Fig.s 12, 13, and 14 illustrate the data organization and prompting approach used for multimodal dialogue tasks.

Fig. 12 illustrates the structure of our multimodal dialogue data format. Each dialogue instance is grounded in a specific 3D scene, denoted by the "scene" field. The "conversations" field contains a sequence of interactions that revolve around this scene, capturing the exchange between the human user and the model. Each turn is marked by a "from" field ("human" or "gpt") and a corresponding "value", which includes multimodal placeholders "<image>" in the first round. An "id" is also assigned to each dialogue instance for indexing purposes.

Fig.s 13 and 14 illustrate the prompt design used in our multimodal dialogue system. Each prompt begins with a system message that sets the context for a conversation between a human and an AI assistant. The user input includes a list of available 3D modality features—such as MultiView, RGBD, BEV, PointCloud, and Voxel—each referenced by a corresponding placeholder token (e.g., `<multiview_dinov2>`). These tokens serve as modality-specific representations rather than raw input data.

Fig. 14 further categorizes prompt formats according to different task types. For dense captioning, the assistant is prompted to describe the appearance and spatial context of a specific object based on its name and ID. For question answering tasks, the user issues natural language queries, optionally with contextual grounding. Visual grounding tasks ask the assistant to return object identifiers that match given textual descriptions. The object IDs used in these tasks are typically generated by an instance segmentation model such as Mask3D [57], ensuring consistent identification across the 3D scene understanding dataset.

F Model Details

F.1 Modality-specific Encoders

Multi-view RGB Encoder. Given an RGB-D video $\boldsymbol{V} \in \mathbb{R}^{V_v \times H \times W \times 3}$, we adopt Maximum Voxel Coverage Sampling (MVCS) (shown in Alg. 1) to select informative keyframes as Multi-view RGB Images $\boldsymbol{I} \in \mathbb{R}^{V \times H \times W \times 3}$, where V_v is the video length, H and W are the image height and width, and V is the predefined number of views. In our implementation, we set V=24. Compared to previous methods [17], our algorithm achieves 100× speed-up in computing coverage by using camera poses instead of depth maps. To focus on task-relevant objects, the algorithm removes low-contribution scene segments (e.g., floor), and voxels located too far from the camera are considered not covered. Finally, for each maximum coverage frame, the algorithm selects its clearest neighbor as the keyframe using Laplacian variance. We use a pre-trained 2D encoder (e.g., DINOv2 [54]) with a trainable modal projector to convert Multi-view RGB Images $\boldsymbol{I}$ into visual tokens $\boldsymbol{F}_{rgb} \in \mathbb{R}^{VN_{rgb} \times D_{rgb}}$, where N_{rgb} is the number of tokens per image, and D_{rgb} is the feature dimension.

Multimodal Dialogue Data Format

```
[
    {
        "id": "0",
        "scene": "scene0000_00",
        "conversations": [
            {
                "from": "human",
                "value": " <image> \nWhat is in the right corner of room by curtains?"
            },
            {
                "from": "gpt",
                "value": "brown cabinet with tv sitting in it"
            },
            {
                "from": "human",
                "value": " <image> \nWhat color table is on the left side of the cabinet?"
            },
            {
                "from": "gpt",
                "value": "light brown"
            }
        ]
    },
    {
        "id": "1",
        "scene": "scene0000_00",
        "conversations": ...
    },
    ...
]
```

Figure 12: Multimodal dialogue data format.

Prompt Template

SYSTEM PROMPT: A chat between a curious human and an artificial intelligence assistant.

The assistant gives helpful, detailed, and polite answers to the human's questions.

USER: The MultiView of the 3D scene is <multiview_dinov2> .

The RGBD of the 3D scene is <rgbd_clip> .

The BEV of the 3D scene is <bev_dinov2> .

The PointCloud of the 3D scene is <pointcloud_pointnet2> .

The Voxel of the 3D scene is <voxel_mask3d> .

What is in the right corner of room by curtains?

ASSISTANT: Brown cabinet with tv sitting in it.

Figure 13: Prompt template.

Multi-view Depth Encoder. Inspired by [33], we leverage depth images $\boldsymbol{D} \in \mathbb{R}^{V \times H \times W \times 1}$ and camera parameters to back-project each 2D patch into 3D space, obtaining its corresponding 3D position. The 2D patch tokens are first extracted from multi-view RGB images using CLIP [55]. These 3D positions are then encoded into 3D embeddings, which are added to 2D patch tokens to form spatially-aware 3D patches. To reduce sequence length while preserving spatial context, we apply a 3D-aware pooling and obtain the final RGB-D features $\boldsymbol{F}_{rgbd} \in \mathbb{R}^{VN_{rgbd} \times D_{rgbd}}$, where N_{rgbd} and D_{rgbd} represent the token count and feature dimension.

BEV Map Encoder. Egocentric images/videos typically lack global scene context, making it difficult for models to understand the overall spatial layout. To address this, we render the 3D mesh in a bird's-eye view (BEV) image

Prompt Template For Different Tasks

SYSTEM PROMPT: A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.

USER: The MultiView of the 3D scene is <multiview_dinov2> .
The RGBD of the 3D scene is <rgbd_clip> .
The BEV of the 3D scene is <bev_dinov2> .
The PointCloud of the 3D scene is <pointcloud_pointnet2> .
The Voxel of the 3D scene is <voxel_mask3d> .

Dense Captioning

Begin by detailing the visual aspects of the <Object_Name> numbered <Object_ID> before delving into its spatial context among other elements within the scene.

ASSISTANT: <Caption>

Visual Question Answering

<Question>

ASSISTANT: <Answer>

Situated Question Answering

<Situation> <Question>

ASSISTANT: <Answer>

Single-object Visual Grounding

What is the ID of the object that matches the description <Description>?

ASSISTANT: <Object_ID>

Multi-object Visual Grounding

Are there objects described as <Description>? Provide the IDs for those objects.

ASSISTANT: <Object_ID>, <Object_ID>

Figure 14: Prompt Template for different tasks.

$\boldsymbol{B} \in \mathbb{R}^{H \times W \times 3}$, where H and W are the height and width, respectively. To enhance object-level understanding, we incorporate instance segmentation into the BEV using numeric labels and colored regions, offering explicit semantic cues. Then, we use DINOv2 [54] to extract BEV features $\boldsymbol{F}_{bev} \in \mathbb{R}^{N_{bev} \times D_{bev}}$, where N_{bev} denotes the number of tokens and D_{bev} is the feature dimension.

Point Cloud Encoder. We apply Farthest Point Sampling (FPS) [56] to the scene-level point cloud to obtain $\boldsymbol{P} \in \mathbb{R}^{N \times C}$, where N is the number of sampled points and C includes the 3D coordinates along with additional attributes such as color, normals, and semantic labels. The sampled points P are then passed through a pre-trained PointNet++ [56] backbone to produce point features $\boldsymbol{F}_{pc} \in \mathbb{R}^{N_{pc} \times D_{pc}}$, where N_{pc} is the number of point tokens and D_{pc} is the feature dimension. In our implementation, we use farthest point sampling (FPS) to acquire 8,192 sampled points ($N_{pc} = 8,192$), where each point contains XYZ coordinates and RGB color attributes ($D_{pc} = 6$).

Voxel Grid Encoder. To extract voxel features, we first voxelize the entire 3D scene and obtain the sparse voxel inputs $\boldsymbol{X} \in \mathbb{R}^{M \times C'}$. Here, M is the number of non-empty voxels and C' is the feature dimension, including sparse tensor coordinates and additional attributes. The voxelized input $\boldsymbol{X}$ is then fed into Mask3D [57], a sparse convolutional U-Net backbone equipped with downsampling and upsampling layers to capture hierarchical context. The voxel-level features are then assigned to their corresponding pre-generated segments, followed by segment-wise average pooling to produce high-level representations $\boldsymbol{F}_{voxel} \in \mathbb{R}^{N_{voxel} \times D_{voxel}}$. Here, N_{voxel} and D_{voxel} are the number and dimension of voxel tokens, respectively.

Subsequently, tokens from five modalities are aligned to the text space via respective adapters: $\boldsymbol{F}'_m = \text{Adapter}_m(\boldsymbol{F}_m) \in \mathbb{R}^{N_m \times D_{txt}}$, where $m \in \{\text{rgb}, \text{rgbd}, \text{bev}, \text{pc}, \text{voxel}\}$ is the modality type and D_{txt} is the target embedding dimension. Finally, the text prompt feature $\boldsymbol{F}_{txt}$, combined with modality-aligned features $\boldsymbol{F}'_m$, composes the unified 3D scene representation: $\boldsymbol{F}_{uni} = \{\boldsymbol{F}_{txt}, \boldsymbol{F}'_{rgb}, \boldsymbol{F}'_{rgbd}, \boldsymbol{F}'_{bev}, \boldsymbol{F}'_{pc}, \boldsymbol{F}'_{voxel}\} \in \mathbb{R}^{N_{uni} \times D_{txt}}$, where N_{uni} is the total token number.

Algorithm 1 Maximum Voxel Coverage Sampling(MVCS) with Voxel Pruning and View Refinement

Require: Scene voxel set V, camera params $\{C_k\}$, budget K, distance limit $d_{\max}$
Ensure: Selected view set S

function SAMPLING(V, $\{C_k\}$, K, $d_{\max}$)
  $V_{\text{scene}} \leftarrow \{\, v \in V \mid v.\text{type} \in \{\text{floor}, \text{ceiling}, \text{wall}\}\}$ ▷ 1. Voxel pruning
  for each view f_k **do**
    $V_k \leftarrow \{\, v \in V_{\text{scene}} \mid \text{visible}(\text{proj}(v, C_k)) \text{ and } \|X_v - X_{C_k}\| \le d_{\max} \,\}$
  end for

  $S \leftarrow \varnothing, \quad U \leftarrow \varnothing$ ▷ 2. Perform greedy selection based on marginal coverage
  while $|S| < K$ **do**
    for each view $f_k \notin S$ **do**
      $g_k \leftarrow |V_k \setminus U|$
    end for
    $f^* \leftarrow \arg\max_k g_k$
    $S \leftarrow S \cup \{f^*\}, \quad U \leftarrow U \cup V_{f^*}$
  end while

  for each index i such that $f_i \in S$ **do** ▷ 3. Selected clear views
    $N \leftarrow \{\, f_j \mid j \in [\max(0, i-2), \min(i+2, n-1)] \,\}$
    for each $f_j \in N$ **do**
      $s_j \leftarrow \text{Var}(\nabla^2 I_{f_j})$
    end for
    $f_{\text{best}} \leftarrow \arg\max_{f_j \in N} s_j$
    $f_i \leftarrow f_{\text{best}}$
  end for
  return S
end function

F.2 Modality-specific Adapters.

To effectively integrate diverse 3D scene representations into the shared embedding space of the language model, we design modality-specific adapters tailored to the characteristics of each input modality. Each adapter employs a lightweight two-layer MLP projection head to map modality-specific embeddings into the unified 4096-dimensional embedding space required by the language backbone, thereby facilitating efficient multimodal feature alignment and fusion.

Multi-view RGB Adapter. The multi-view RGB adapter processes concatenated multi-view RGB features, typically aggregating visual details from multiple camera views (e.g., 8 views), resulting in a 12288-dimensional input. It projects these into the common embedding space via: Linear$(12288 \to 4096) \to$ GELU $\to$ Linear$(4096 \to 4096) \to$ LayerNorm. This structure effectively reduces dimensional redundancy from multiple views and normalizes feature distributions for stable integration.

RGBD Adapter. The RGBD adapter operates on 1024-dimensional spatially-aware RGBD embeddings, leveraging depth-enhanced RGB features. It applies: Linear$(1024 \to 4096) \to$ GELU $\to$ Linear$(4096 \to 4096)$, to unify depth-aware visual cues with the broader modality embedding space.

BEV Adapter. The BEV adapter receives BEV-encoded features from `BEVDinov2Encoder`, which inherently capture spatial structures from a top-down viewpoint in 1536-dimensional embeddings. It aligns them via: Linear$(1536 \to 4096) \to$ GELU $\to$ Linear$(4096 \to 4096) \to$ LayerNorm. This ensures consistent spatial semantic representations across modalities.

Point Cloud Adapter. The point cloud adapter adapts the sparse geometric features extracted by the `PointNet2SegEncoder`, which generates a compact 256-dimensional representation from raw point cloud data. It expands and aligns these features using: Linear$(256 \to 4096) \to$ GELU $\to$ Linear$(4096 \to 4096) \to$ LayerNorm, preserving critical geometric semantics for downstream tasks.

Voxel Adapter. The voxel adapter uniquely incorporates five parallel linear branches tailored to voxel features of varying dimensionalities (256, 128, or 96), accommodating voxel grids captured at multiple spatial resolutions. Each branch independently projects features via: Linear $\to$ LayerNorm $\to$ GELU $\to$ Dropout $\to$ Linear $\to$ LayerNorm, producing uniform 4096-dimensional embeddings to robustly represent volumetric structural information across scales.

By individually tailoring adapter structures to the intrinsic properties of each input modality, these designs collectively ensure robust alignment and effective integration of heterogeneous sensor inputs within the language model.

Low-Rank Adaptation (LoRA) in Stage I. In Stage I training, we utilize Low-Rank Adaptation (LoRA) to efficiently fine-tune the LLM backbone, minimizing the number of trainable parameters. Each LoRA module is characterized by a rank of 32, a scaling factor (α) of 64, and a dropout rate of 0.05. LoRA modules do not include bias terms, further streamlining the adaptation process. This low-rank structure substantially reduces computational overhead, enabling stable and effective fine-tuning while maintaining the representational capacity required to integrate diverse multimodal information.

Mixture of Expert in Stage II. In Stage II, we strategically integrate sparse Mixture-of-Experts (MoE) layers into selected transformer blocks ([8, 12, 16, 20, 24, 28]) within the language backbone. Each MoE layer consists of 8 parallel expert modules, implemented as specialized LLaMA-style multilayer perceptrons (MLP). These experts expand the feature dimension from 4096 to an intermediate dimension of 11008 via parallel gated projections (gate_proj and up_proj), followed by SiLU activations and dimensional reduction back to 4096 (down_proj). Tokens are dynamically routed to these experts using a learnable Top-K gating network, which adaptively selects the most suitable experts based on token-level semantic characteristics. In our implementation, we set K to 2. This design encourages sparsity and computational efficiency, enabling effective modeling of heterogeneous modality information while preserving scalability. The adaptive routing mechanism thus enhances the model's ability to exploit specialized knowledge, significantly improving performance across diverse and complex multimodal 3D scene understanding tasks.