

Inst3D-LMM: Instance-Aware 3D Scene Understanding with Multi-modal Instruction Tuning

Hanxun Yu^{1*}, Wentong Li^{2*}, Song Wang¹, Junbo Chen^{3†}, Jianke Zhu^{1†}

¹Zhejiang University ²Nanjing University of Aeronautics and Astronautics ³Udeer.ai

{hanxun.yu, songw, jkzhu}@zju.edu.cn, wentong.li@nuaa.edu.cn, junbo@udeer.ai

Abstract

*Despite encouraging progress in 3D scene understanding, it remains challenging to develop an effective Large Multi-modal Model (LMM) that is capable of understanding and reasoning in complex 3D environments. Most previous methods typically encode 3D point and 2D image features separately, neglecting interactions between 2D semantics and 3D object properties, as well as the spatial relationships within the 3D environment. This limitation not only hinders comprehensive representations of 3D scene, but also compromises training and inference efficiency. To address these challenges, we propose a unified **Instance-aware 3D Large Multi-modal Model (Inst3D-LMM)** to deal with multiple 3D scene understanding tasks simultaneously. To obtain the fine-grained instance-level visual tokens, we first introduce a novel Multi-view Cross-Modal Fusion (MCMF) module to inject the multi-view 2D semantics into their corresponding 3D geometric features. For scene-level relation-aware tokens, we further present a 3D Instance Spatial Relation (3D-ISR) module to capture the intricate pairwise spatial relationships among objects. Additionally, we perform end-to-end multi-task instruction tuning simultaneously without the subsequent task-specific fine-tuning. Extensive experiments demonstrate that our approach outperforms the state-of-the-art methods across 3D scene understanding, reasoning and grounding tasks. Source code is available at: <https://github.com/hanxunyu/Inst3D-LMM>.*

1. Introduction

Building Large Multi-modal Models (LMMs) for 3D scene understanding becomes an emerging research topic with significant potential for advancing autonomous robotics [37]. For example, the interactive embodied agents [40] are expected to interpret 3D layouts and predict

object locations based on human instructions.

Traditional 3D scene understanding methods [31, 47] are typically tailored for individual downstream tasks, such as 3D Visual Grounding (3D-VG), 3D Question Answering (3D-QA) and 3D Dense Captioning (3D-DC). In contrast, LMMs are able to handle various 3D perception tasks within a single model. Some methods [16, 44] primarily focus on translating 3D points into the space of 2D Vision Language Models (VLMs) or directly leveraging multi-view 2D features as 3D representations. Alternatively, other approaches [6, 46, 60] directly encode the features of 3D points and facilitate the alignment with LLM using 3D-text instruction data. However, they often require multi-stage alignment or language-scene pre-training, complicating the development of a versatile model capable of handling multiple tasks. To enable a unified 3D LMM framework, recent work [19] decomposes the input 3D scene into a set of individual object proposals, each identified by unique tokens to capture instance-level 3D object features explicitly. While this approach exhibits promising results, it neglects the interactions between the 2D semantic features and the properties of 3D objects, as well as the spatial relationship modeling among objects in 3D environments. This oversight further results in substantial token costs for the LLM, thereby hindering both training and inference efficiency.

In this paper, we propose Inst3D-LMM, an effective **Instance-aware 3D Large Multi-modal Model** that tackles multiple 3D-language tasks without resorting to task-specific fine-tuning. Our approach fully leverages the powerful 2D Vision Foundation Models (VFM) and 3D specialist models to extract enriched 2D and 3D features at the instance level respectively. As shown in Figure 1, in contrast to previous methods, our approach is able to generate fine-grained instance-level representations that encapsulate both geometric and semantic properties, and scene-level representations that capture intricate pairwise spatial relationships among objects in a 3D scene. Moreover, our method results in minor token costs for the LLM, thereby enhancing both training and inference efficiency. By leveraging this instance-aware methodology, our Inst3D-LMM

*Equal contribution.

†Corresponding authors.

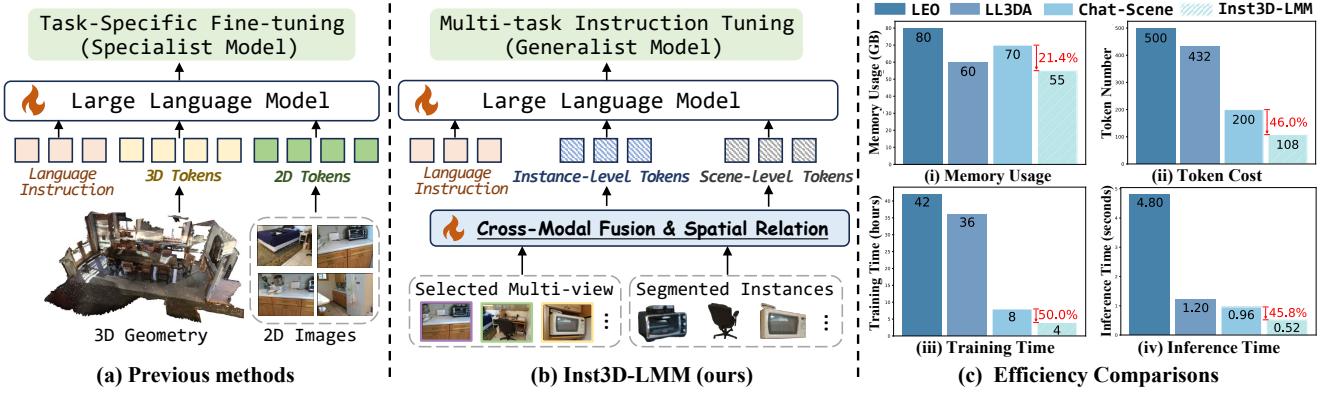


Figure 1. Comparisons between previous 3D LMMs and our proposed Inst3D-LMM. (a) Previous methods [6, 16, 19, 20] typically encode the features of 3D points or 2D images separately and concatenate them directly, often requiring task-specific fine-tuning for different tasks. (b) Illustration of Inst3D-LMM. Our method integrates 2D/3D cross-modal information and captures the intricate spatial relations among objects within 3D environments to generate tight but informative instance/scene-level tokens for the LMM. (c) Compared with other 3D LMMs, our Inst3D-LMM requires fewer computational resources, while offering faster training and inference speeds.

significantly improves the LLM’s ability to comprehend 3D scenes with respect to both efficiency and accuracy.

Specifically, we introduce a novel Multi-view Cross-Modal Fusion (MCMF) module that effectively infuses enriched multi-view 2D features into the original 3D features with coarse semantics. A learnable [CLS] token is introduced to aggregate the characteristics of each 2D view, enabling efficient multi-view 2D-to-3D cross-modal transformation. To capture intricate pairwise spatial relationships among objects in a 3D scene, we then propose a 3D Instance Spatial Relation (3D-ISR) module. A spatial condition self-attention between manifold position embeddings and instance-level tokens is presented to produce relation-aware scene-level representations. The resulting instance-level and scene-level representations are subsequently fed into the LLM for end-to-end multi-task instruction tuning.

Extensive experiments across various tasks, including 3D-VG, 3D-QA and 3D-DC, demonstrate that our approach outperforms previous state-of-the-art methods with leading 3D scene understanding, grounding and reasoning capabilities. Unlike most existing methods that focus on close-set scene understanding or require per-task fine-tuning, our Inst3D-LMM operates as a generalist model. We believe this work lays a fundamental step towards unifying diverse 3D vision-language tasks in generative language modeling.

To summarize, our contributions are as follows:

- We propose a unified and efficient instance-aware LLM-based framework, called Inst3D-LMM for various 3D scene understanding tasks with end-to-end multi-modal instruction tuning. Serving as a generalist model, our approach demonstrates superior performance across 3D scene understanding, reasoning and spatial localization.
- We utilize 2D VFM to extract multi-view contextual features for each 3D instance and then devise a Multi-view Cross-Modal Fusion (MCMF) module to effectively en-

hance instance-level feature representations by jointly integrating 3D geometry and 2D semantic priors.

- A 3D Instance Spatial Relation (3D-ISR) module is introduced to boost the capability of LMM in understanding the complex spatial details within 3D scenes.

2. Related Works

3D Scene Understanding with Language. In 3D scene understanding, there is a surge of interest in making use of language queries to capture user intentions for various downstream tasks, such as 3D Visual Grounding [5, 45], 3D Question Answering [31, 34] and 3D Dense Captioning [8, 22]. Specifically, 3D Visual Grounding entails localizing target objects based on language queries. Moreover, 3D Question Answering demands robust 3D spatial perception and reasoning. 3D Dense Captioning involves localizing and describing objects in 3D scenes. The conventional methods typically focus on a specific task. Instead, 3D visual grounding and dense captioning tasks are combined by leveraging their complementary aspects [2, 9, 32, 48]. Recent efforts like 3D-VLP [23] and 3D-VisTA [62] attempt to establish a universal framework by pre-aligning 3D scenes with their corresponding textual descriptions. In contrast to our Inst3D-LMM, most existing methods still focus on close-set scene understanding, which requires either task-specific fine-tuning or striving to build specialized models.

3D Large Multi-modal Models. Inspired by the significant advancements in Large Language Models (LLMs), researchers extend LLM’s knowledge to encompass 3D modality [30, 50, 59, 61]. Point-LLM [14] and Imagebind-LLM [15] have succeeded in bridging the gap between 3D visuals and text by utilizing extensive 3D object datasets. However, these models struggle with interpreting complex spatial relationships in 3D scenes. Another promising di-

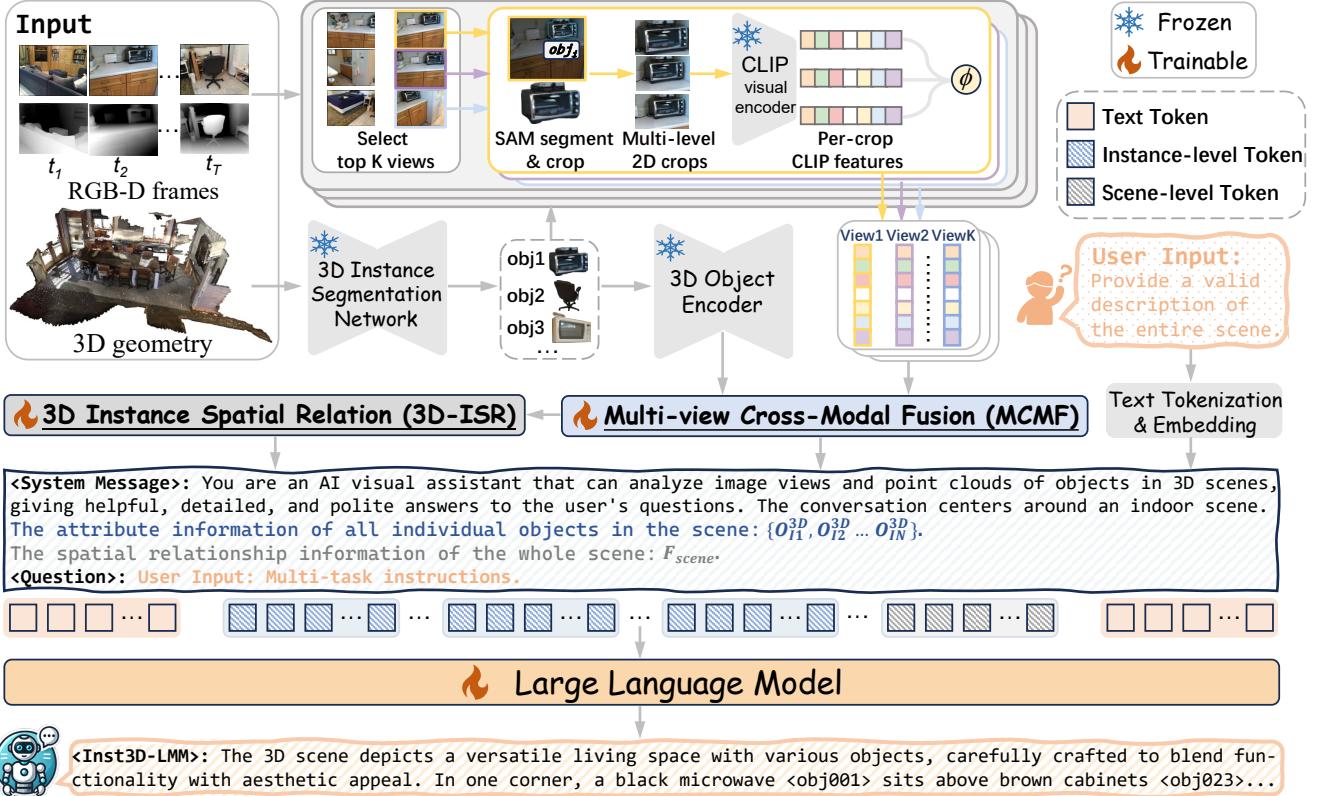


Figure 2. **Overview of our proposed Inst3D-LMM.** Our pipeline takes as input point clouds of a 3D indoor scene, along with RGB-D images. We first employ the pre-trained 3D specialist models and 2D VFM to extract 3D proposals and multi-view 2D semantic features, respectively. We then suggest the MCMF module to generate fine-grained instance-level tokens. A 3D-ISR module is further introduced to create relation-aware scene-level tokens based on spatial distances. By leveraging the constructed 3D-language prompts, we conduct multi-task instruction tuning to simultaneously handle various 3D tasks.

rection focuses on developing the scene-level 3D LMMs. Hong et al. [16] encodes projected 3D features using a 2D vision encoder and incorporates location tokens to augment LLM vocabularies. Other methods [6, 46, 60] directly encode point clouds and utilize 3D scene-text data for better visual interaction through pre-alignment. Huang et al. [18] employs a three-stage training scheme and adopts object identifiers to learn individual object attributes. Chen et al. [7] leverages special referent tokens for precise referencing and grounding. Wang et al. [44] encodes point clouds and RGB images separately. However, some crucial semantic pixels are often lost during their sparse fusion processes, resulting in a coarse visual semantic representation. In this work, we propose an effective instance-aware framework to fuse fine-grained cross-modal information and encode spatial relations, which achieves promising results on multiple 3D-language tasks as a generalist model.

3. Methodology

Our goal is to enable LLM to understand the 3D environment and perform various visual interaction tasks based

on human instructions. Figure 2 illustrates the architecture of our framework. In this section, we first introduce how to extract features at the instance level using pre-trained 3D models [39, 59] and 2D VFM [24, 36], respectively. Secondly, we present a novel Multi-view Cross-Modal Fusion (MCMF) module to obtain fine-grained instance-level tokens, which is specially designed to effectively integrate 3D geometric features with their corresponding multi-view 2D semantic features. Thirdly, we introduce the 3D Instance Spatial Relation (3D-ISR) module to enhance LLM’s ability to capture spatial information at the scene level, which generates relation-aware tokens through attention-based analysis of spatial relationships among different 3D proposals. Under our proposed framework, the MCMF and 3D-ISR modules are jointly optimized, enabling mutual enhancement. Finally, we conduct end-to-end multi-task instruction tuning to address a range of 3D scene understanding tasks simultaneously.

3.1. Instance-Level Feature Extraction

3D Feature Extraction. We first segment the 3D point clouds of each individual instance in a *class-agnostic* man-

ner by leveraging a pre-trained 3D instance segmentation model [39]. We only retain the predicted binary 3D instance masks while ignoring their closed-vocabulary class labels. All instance proposals in one scene are represented by $O^{3D} = (O_1^{3D}, \dots, O_N^{3D})$, where $O_i^{3D} = [\text{coordinate}, \text{color}]$ consists of the attributes of each instance. We obtain all instance proposals of a scene, and the pre-trained 3D encoder E_o [59] is used to extract their instance-level features, *i.e.*, $f_o^{3D} = E_o(O^{3D})$.

2D Feature Extraction. Due to the inherent sparsity of 3D point cloud data, the previous methods have difficulties in generating discriminative features of each object. In this work, we employ powerful 2D VFM to extract 2D semantic features for each 3D instance. As in [3, 41], we firstly project the point cloud of each instance O_i^{3D} onto the image plane and then select the top K views according to the number of visible points. To obtain the accurate 2D masks, we randomly sample $k_{\text{sample}} = 5$ points as the input prompts for SAM [24] to deal with noisy bounding boxes with outliers. Hereby, we select the high-quality mask with the highest confidence score. To enrich features with contextual information, these masks are employed to generate multi-level (L) crops of the selected images that are further fed into the pre-trained CLIP vision encoder [36] to extract features with language-aligned embedding space. Finally, we aggregate multi-level features of sub-images from the same frame to form the 2D multi-view features O^{2D} .

3.2. Multi-view Cross-Modal Fusion

To better fuse 3D geometry priors with 2D multi-view semantic priors, we introduce an effective Multi-view Cross-Modal Fusion (MCMF) module that generates the enriched token representations for each 3D instance before being fed into LLM. The architecture of MCMF is crafted with a coarse-to-fine framework, as shown in Figure 3. To map 3D object features $O^{3D} \in \mathbb{R}^{1 \times N \times D^{3d}}$ and 2D CLIP features $O^{2D} \in \mathbb{R}^{K \times N \times D^{2d}}$ into the embedding space of LLM with the dimensionality of D , we utilize a simple two-layer MLP with a LayerNorm and GELU in between, yielding $O^{3D'}$ and $O^{2D'}$, respectively. N represents the number of instances, and K is the number of images for each object.

Subsequently, we introduce a Cross-Modal Injection Block to transform the enriched semantic priors from 2D multi-view representations into 3D instance features. Moreover, we adopt a straightforward self-attention layer to further enhance such 3D instance features. For 2D multi-view features, we append a learnable [CLS] token t_k to the flattened CLIP feature maps in order to adaptively encapsulate the global semantic representation of the k -th view. Then, we apply a self-attention layer to multi-view features, enabling semantic gathering to derive t'_k for k -th view as:

$$t'_k = \phi(\text{SelfAttn}([t_k, O_{\text{view}, k}^{2D'}])), \quad (1)$$

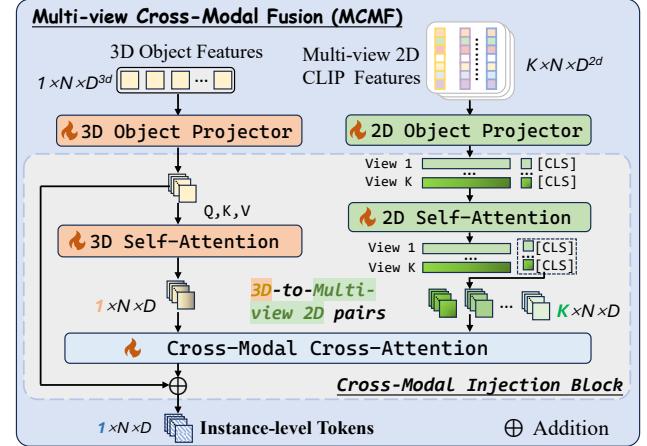


Figure 3. Architecture of the proposed Multi-view Cross-Modal Fusion (MCMF) module.

where $1 \leq k \leq K$, $O_{\text{view}, k}^{2D'} \in \mathbb{R}^{N \times D}$. ϕ is an indicator function that outputs the updated [CLS] token as the first one from the token list.

We obtain the processed 3D token embedding $O^{3D'} \in \mathbb{R}^{1 \times N \times D}$ containing coarse semantics, where each instance in $O^{3D'}$ corresponds to multiple views of the processed 2D CLIP features $O^{2D'}$. It can be represented as follows:

$$O_i^{2D'} = [t'_1, \dots, t'_K], \quad i \in [1, N]. \quad (2)$$

To infuse the enriched semantic 2D features into 3D geometric feature, we construct the 3D-to-2D-Multi-view pairs, *i.e.* each 3D instance feature in $O^{3D'} \in \mathbb{R}^{1 \times N \times D}$ corresponds to 2D features across K views in $O^{2D'} \in \mathbb{R}^{K \times N \times D}$. In particular, we utilize the preliminary $O^{3D'}$ as queries, while the 2D multi-view visual features $O^{2D'}$ as enriched reference keys and values. Inspired by [26], the injection process is conducted via a cross-attention layer, which encourages 3D instance queries to absorb the fine-grained semantics of keys and values. This results in the enhanced 3D instance features $O_f^{3D} \in \mathbb{R}^{1 \times N \times D}$ as below:

$$O_f^{3D} = \text{CrossAttn}(O^{3D'}, O^{2D'}). \quad (3)$$

Furthermore, we leverage a residual operation by adding the input $O^{3D'}$ to retain the basic characteristics in generating the instance-level 3D visual tokens, $O_I^{3D} = O_f^{3D} + O^{3D'}$.

3.3. 3D Instance Spatial Relation

Motivated by [5, 28, 62], we develop an effective 3D Instance Spatial Relation (3D-ISR) module to boost LLM's capabilities in assimilating spatial information within the 3D scene, as shown in Figure 4. 3D-ISR utilizes the instance-level 3D visual tokens $\{O_{I1}^{3D}, O_{I2}^{3D}, \dots, O_{IN}^{3D}\}$ derived from MCMF module along with the corresponding center coordinates $\{C_1, C_2, \dots, C_N\}$ of all instances as inputs. For the i -th instance, we define its center coordinates

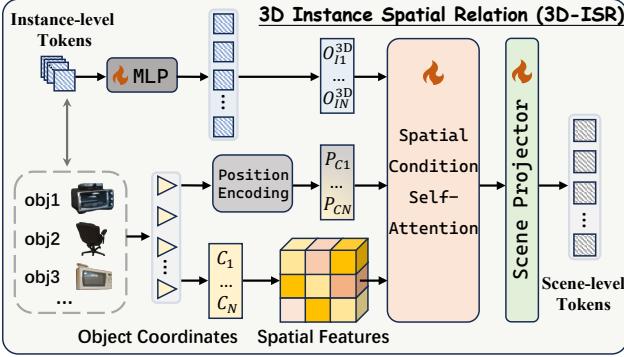


Figure 4. Illustration of the 3D Instance Spatial Relation (3D-ISR) module in our framework.

as $C_i = (x_i, y_i, z_i)$. For each pair of instance-level tokens $\{O_{Ii}^{3D}, O_{Ij}^{3D}\}$, we calculate their Euclidean distance $d_{ij} = \|C_i - C_j\|_2$ as well as their horizontal angle θ_h and vertical angle θ_v . Specially, $\theta_h = \arctan 2((y_j - y_i)/(x_j - x_i))$ and $\theta_v = \arcsin((z_j - z_i)/d_{ij})$. By making use of these positional parameters, we generate the pairwise spatial features $S = \{s_{ij}\} \in \mathbb{R}^{N \times N \times 5}$ as:

$$s_{ij} = [\sin(\theta_h), \cos(\theta_h), \sin(\theta_v), \cos(\theta_v), d_{ij}] . \quad (4)$$

Inspired by language-conditioned self-attention presented in ViL3DRel [5], we suggest a spatial-conditioned self-attention module. Specially, we create position embeddings \mathcal{P} via the absolute positional encoding (PE), i.e. $\mathcal{P} = \text{PE}(C_1, C_2, \dots, C_N)$. A spatial conditioned attention weight l_i is computed to select the relevant spatial relations for each instance O_{Ii}^{3D} , which is formulated as below:

$$l_i = W_P^\top (\mathcal{P}_i + O_{Ii}^{3D}), \quad (5)$$

where $W_P \in \mathbb{R}^{D \times 5}$ is a learnable parameter and the bias term is omitted for simplicity. The spatial-conditioned attention map ω_{ij} is computed by combining s_{ij} , l_i and l_j , i.e. $\omega_{ij} = l_i \cdot s_{ij} \cdot l_j$.

Through this positional transformation, the spatial-conditioned attention map ω_{ij} encapsulates the pairwise spatial relationship across the entire 3D scene for each instance. We further integrate the attention map with instance-level visual tokens, which is formulated as below:

$$F_i = \sum_{j=1}^N \omega_{ij} O_{Ij}^{3D}. \quad (6)$$

Finally, the instance-level representations are concatenated and further processed by a transformer-based encoder Γ , followed by a max-pooling layer (Pool) and a simple two-layer MLP. The scene-level relation-aware tokens $F_{\text{scene}} \in \mathbb{R}^D$ are generated for the entire 3D scene as follows:

$$F_{\text{scene}} = \text{MLP}(\text{Pool}([\Gamma([F_1, F_2, \dots, F_N]))]). \quad (7)$$

Thus, we obtain the final token representations for LLM processing, i.e. instance-level visual tokens O_{Ii}^{3D} and spatial relation-aware scene-level tokens F_{scene} , respectively.

3.4. End-to-End Multi-task Instruction Tuning

3D-Language Prompts. Given an LMM, clear and explicit system messages and instructions are essential to support a range of downstream 3D vision-language tasks. For different tasks, we adopt various task-specific instruction templates with 3D features to generate uniform instruction data, enabling multi-task training. *Please refer to the Supplementary Material for more detailed information.*

Instruction Tuning. We conduct end-to-end multi-task fine-tuning to fully leverage the capabilities of our Inst3D-LMM framework. The MCMF and 3D-ISR modules enable the LLM with robust 3D scene understanding, grounding and reasoning abilities. We freeze the pre-trained 3D object encoder while updating both MCMF and 3D-ISR modules, as well as LLM. Upon completion of end-to-end multi-task instruction tuning, Inst3D-LMM can effectively handle various 3D vision-language tasks simultaneously, without fine-tuning on specific tasks, as illustrated in Figure 7.

4. Experiments

4.1. Experimental Setting

Datasets and Benchmarks. In this work, we conduct our experiments on the ScanNetv2 dataset [11], an extensive indoor 3D scene dataset comprising 1,513 scenes. This dataset includes 3D point clouds, RGB-D frames, and detailed point-level instance segmentation annotations. The whole dataset is divided into 1,201 scenes for training and 312 scenes for validation, with all subsequent benchmarks adhering to these training/validation splits. Our evaluation encompasses a range of 3D scene understanding benchmarks, including ScanRefer [4] and Multi3DRefer [57] for single- and multi-object 3D Visual Grounding, respectively, ScanQA [1] for 3D Question Answering, and Scan2Cap [8] for 3D Dense Captioning. These datasets are converted into a uniform instruction format for multi-task instruction tuning and performance assessment.

Implementation Details. For 3D feature extraction, we leverage the 3D instance segmentation model Mask3D [39] pre-trained on ScanNet200 dataset [38], alongside the 3D object encoder Ulip2 [49]/Uni3D [59] based on ViT-L/14 [13]. For 2D feature extraction, we adopt the ViT-H-based SAM [24] to obtain high-quality masks. Moreover, we extract 2D semantic features using the vision encoder from CLIP-ViT-L/14-336px [36]. These pre-trained models are kept frozen. In our method, we apply multi-view selection and multi-level crops to 2D images, setting $K = 5$ views and $L = 3$ levels. We use Vicuna1.5-7B [10] as our basic LLM, which is fine-tuned from LLaMA2 [42].

Method	Overall		Unique		Multiple		Multi3DRefer	
	Acc@0.25↑	Acc@0.50↑	Acc@0.25↑	Acc@0.50↑	Acc@0.25↑	Acc@0.50↑	F1@0.25↑	F1@0.50↑
Closed-set, full sup.								
ScanRefer [4]	37.3	24.3	65.0	43.3	30.6	19.8	—	—
3DVG-Trans [58]	45.9	34.5	77.2	58.5	38.4	28.7	30.2	25.5
M3DRef-CLIP [57]	51.9	44.7	—	77.2	—	36.8	42.8	38.4
ConcreteNet [43]	56.1	49.5	86.1	79.2	47.5	40.9	—	—
CORE-3DVG [52]	56.8	43.8	85.0	67.1	51.8	39.8	—	—
Zero-shot								
LLM-Grounder [51]	17.1	5.3	30.8	22.6	16.3	12.9	—	—
Visual-Programming [55]	36.4	32.7	<u>63.8</u>	<u>58.4</u>	<u>27.7</u>	<u>24.6</u>	—	—
Specialist								
OpenScene [35]	13.0	5.1	20.1	13.1	11.1	4.4	—	—
3D-LLM(Flamingo) [16]	21.2	—	—	—	—	—	—	—
3D-LLM(BLIP2-flant5) [16]	30.3	—	—	—	—	—	—	—
Chat-3D v2 [18]	35.9	30.4	61.2	57.6	25.2	22.6	45.1	41.6
ReGround3D [60]	53.1	41.1	—	—	—	—	—	—
Generalist								
LAMM [53]	—	3.4	—	—	—	—	—	—
3DMIT [27]	10.7	7.2	—	—	—	—	—	—
Grounded 3D-LLM [7]	47.9	44.1	—	—	—	—	45.2	40.6
Chat-Scene [19]	<u>55.5</u>	<u>50.2</u>	—	—	—	—	<u>57.1</u>	<u>52.4</u>
Inst3D-LMM	57.8	51.6	88.6	81.5	48.7	43.2	58.3	53.5

Table 1. Quantitative results for 3D Visual Grounding on ScanRefer and Multi3DRefer validation sets. In the ScanRefer dataset, scenes are labeled as “unique” (one object per class) or “multiple” (more than one). Closed-set methods are fully supervised for specific datasets. “Zero-shot” refers to methods that directly use LLMs without fine-tuning. “Specialist” and “Generalist” categorize methods fine-tuned for specific tasks versus those trained jointly. **Bold** and underlined numbers indicate the best and the second-best results, respectively.

Method	# 3D Data for Alignment	ScanQA						Scan2Cap@0.50			
		B-1↑	B-4↑	METEOR↑	ROUGE↑	CIDER↑	EM↑	B-4↑	METEOR↑	ROUGE↑	CIDER↑
Closed-set, full sup.											
VoteNet [12] + MCAN [54]	—	28.0	6.2	11.4	29.8	54.7	17.3	—	—	—	—
ScanRefer [4] + MCAN [54]	—	26.9	7.9	11.5	30.0	55.4	18.6	—	—	—	—
ScanQA [8]	—	30.2	10.1	13.1	33.3	64.9	21.0	—	—	—	—
Scan2Cap [1]	—	—	—	—	—	—	—	22.4	21.4	43.5	35.2
3D-VisTA [62]	—	34.2	13.1	15.2	38.6	76.6	27.0	34.0	27.1	54.3	66.9
LLM-based Methods											
LLaVA (zero-shot) [29]	—	7.1	0.3	10.5	12.3	5.7	0.2	1.5	8.3	19.6	3.2
LAMM [53]	25K	26.8	5.8	10.0	23.6	42.4	9.8	—	—	—	—
3D-LLM(Flamingo) [16]	675K	30.3	7.2	12.2	32.3	59.2	20.4	5.9	11.4	29.9	—
3D-LLM(BLIP2-flant5) [16]	675K	<u>39.3</u>	12.0	14.5	35.7	69.4	20.5	8.1	13.1	33.2	—
Chat-3D v2 [18]	38K	38.4	7.3	<u>16.1</u>	<u>40.1</u>	77.1	<u>21.1</u>	31.8	22.3	50.2	63.9
LL3DA [6]	38K	—	13.3	15.4	37.0	75.7	—	35.9	<u>25.6</u>	<u>54.6</u>	65.2
Grounded 3D-LLM [7]	107K	—	13.4	—	—	72.7	—	35.5	—	—	70.6
Chat-Scene [19]	38K	—	<u>14.3</u>	—	—	87.7	—	<u>36.3</u>	—	—	<u>77.1</u>
Inst3D-LMM	38K	43.5	14.9	18.4	42.6	88.6	24.6	38.3	27.5	57.2	79.7

Table 2. Quantitative results for 3D Question Answering and 3D Dense Captioning on the ScanQA and Scan2Cap datasets.

Our fine-tuning process utilizes LoRA [17]. We adopt the AdamW optimizer with a weight decay of 0.02. All experiments are conducted on 8 NVIDIA A100 GPUs.

4.2. Main Results

3D Visual Grounding. We first report the visual grounding performance on ScanRefer and Multi3DRefer validation datasets. As shown in Table 1, our approach outperforms the state-of-the-art model, Chat-Scene [19], by **+2.3%** Acc@0.25 and **+1.2%** F1@0.25 on ScanRefer and

Multi3DRefer, respectively. Compared to Specialist and closed-set methods, Inst3D-LMM, *trained with a generalist approach*, achieves competitive performance. Figure 5 displays typical visual comparison results.

3D Question Answering. We then compare Inst3D-LMM with previous leading methods on the ScanQA validation set. Table 2 reports the results. Our Inst3D-LMM consistently outperforms these methods, including the recent LL3DA [6], Grounded 3D-LLM [7] and Chat-Scene [19].

3D Dense Captioning. This task involves the localization

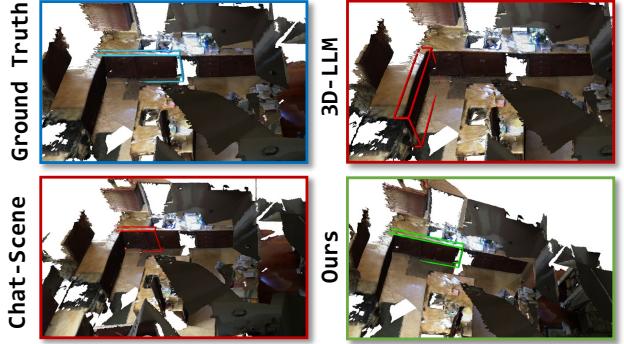


Figure 5. Visual comparisons in 3D Visual Grounding. The rendered images show the ground-truth (blue), incorrectly identified objects (red), and correctly identified objects (green). The colored text indicates the results of text decoupling.

and description of instances. As shown in Table 2, our approach achieves 38.3% B-4@0.50 and 79.7% C@0.50 on Scan2Cap, which exceeds the closed-set expert model 3D-VisTA [62] by +4.3% and +12.8%, and still outperforms LLM-based method LL3DA by +2.4% and +14.5%.

4.3. Ablation Study

In this section, we perform ablation experiments to thoroughly evaluate the effectiveness of components. *Please refer to the Supplementary Material for more analysis.*

Effects of Multi-view Cross-Modal Fusion (MCMF). The MCMF module aims to augment LLM’s understanding of 3D instances’ geometric and semantic attributes. We explore several methods to combine these features, including direct concatenation for projection (‘Concat.’), parallel projection followed by combination (‘Parallel’), and vanilla cross-attention (‘Cross-Attention’). To encode the basic spatial information, we concatenate the results of absolute positional encoding (PE) with the 3D features. As shown in Table 3, MCMF outperforms all these methods on ScanRefer, ScanQA and Scan2Cap datasets. We also observe that the performance of 3D LMM is significantly improved by incorporating multi-view 2D CLIP information compared to using only 3D geometric features. These results demonstrate the effectiveness of our MCMF approach.

Method	ScanRefer		ScanQA		Scan2Cap@0.50	
	Acc@0.25↑	Acc@0.50↑	B-1↑	CIDE↑	B-4↑	CIDE↑
w/o Multi-view 2D	36.0	31.9	31.5	57.3	19.9	54.3
Concat.	38.8	34.4	36.6	65.4	25.2	63.3
Parallel	37.6	33.5	35.3	63.9	24.8	62.1
Cross-Attention	39.2	36.5	37.4	66.1	25.5	64.3
MCMF	46.7	41.9	41.5	78.6	32.7	68.2

Table 3. Ablation evaluations of the proposed Multi-view Cross-Modal Fusion (MCMF) module.

Impacts of 3D Instance Spatial Relation (3D-ISR). We further investigate 3D-ISR module to analyze its impact on

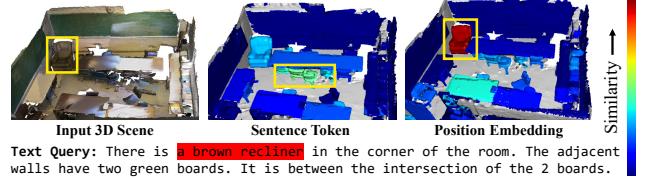


Figure 6. A visualization of the similarity score between text query and segmented 3D proposals. We compare sentence tokens used in ViL3DRel [5] with position embeddings employed in 3D-ISR.

3D spatial understanding using only 3D geometry priors. We compare 3D-ISR against previous spatial relation modeling methods, such as the 3D localization mechanism in ViL3DRel [5] and 3D-LLM [16], and relation-aware token generation in Chat-3D-v2 [18]. Table 4 reports the comparison results. It can be observed that our 3D-ISR consistently surpasses these approaches. Specifically, 3D-ISR significantly enhances performance on grounding tasks (ScanRefer and Multi3dRefer). This aligns with the motivation of our design. We also provide a visualization, as shown in Figure 6, of the similarity between the text query and segmented 3D proposals after processing by LLM. Compared to ViL3DRel [5] using sentence tokens, our 3D-ISR more accurately captures spatial relationships within the scene by employing position embeddings of objects.

Method	ScanRefer		Multi3dRefer		ScanQA	
	Acc@0.25↑	Acc@0.50↑	F1@0.25↑	F1@0.50↑	B-1↑	CIDE↑
w/o spatial relation	35.6	29.8	34.8	27.3	29.2	55.6
ViL3DRel [5]	36.1	30.5	38.4	35.5	32.8	58.9
3D-LLM [16]	39.2	36.8	42.4	35.2	32.5	60.2
Chat-3D-v2 [18]	40.8	37.5	41.6	37.8	35.1	64.6
3D-ISR	48.3	44.1	46.2	41.3	39.1	72.3

Table 4. Ablation study to verify the effectiveness of our 3D Instance Spatial Relation (3D-ISR) module.

Ablations within MCMF and 3D-ISR modules. We also explore the effectiveness of key designs in MCMF and 3D-ISR modules. The learnable [CLS] token in the MCMF module is introduced to aggregate 2D multi-view features. We compare it against alternative methods, such as token max pooling and Q-Former [25]. Table 5 reports the comparison results, demonstrating the efficacy of the learnable [CLS] token design. Besides, to assess the impact of different pairwise spatial features in the 3D-ISR module, we evaluate the model using only distance or orientation information to compute spatial features in Eq 4. Results in Table 6 reveal that distance has a greater impact on the model’s grounding capability, while orientation is more crucial in handling Q&A tasks. Combining both pairwise distance and orientation yields the best overall performance.

Mutual Benefits of MCMF and 3D-ISR. As shown in Table 7, the model integrating both MCMF and 3D-ISR consistently outperforms those utilizing either module alone across multiple tasks. To further verify its effectiveness, we combine MCMF with absolute positional encoding (PE) for the basic spatial relation, and directly adopt feature con-

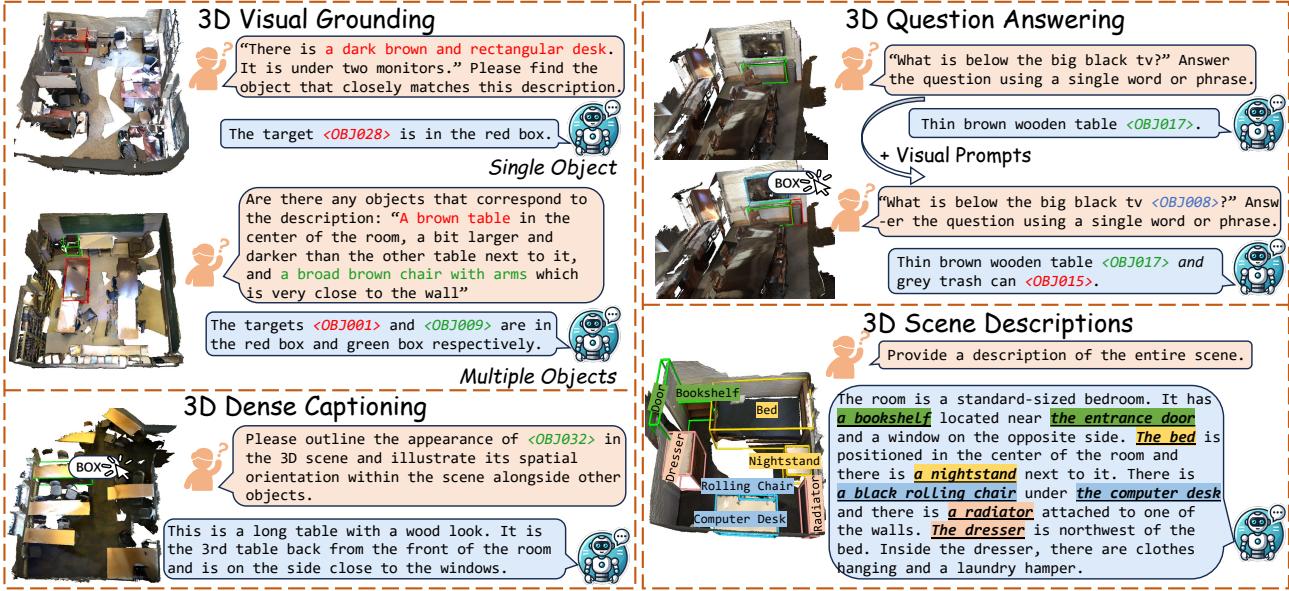


Figure 7. Qualitative illustration of Inst3D-LMM across various 3D-language tasks in diverse 3D environments.

Method	ScanRefer		ScanQA		Scan2Cap@.050	
	Acc@0.25↑	Acc@0.50↑	B-1↑	CIDER↑	B-4↑	CIDER↑
Max Pooling	48.7	44.6	39.2	74.8	30.6	65.5
Q-Former [25]	53.2	47.9	44.3	87.9	35.4	78.6
CLS Token	57.8	51.6	43.5	88.6	38.3	79.7

Table 5. Ablation study of learnable CLS token in MCMF module.

Method	ScanRefer		Multi3DRefer		ScanQA	
	Acc@0.25↑	Acc@0.50↑	F1@0.25↑	F1@0.50↑	B-1↑	CIDER↑
w/ Dist (only)	43.2	38.0	42.4	37.8	31.2	65.7
w/ Ort (only)	39.7	33.6	40.5	35.9	32.5	66.8
Dist + Ort	48.3	44.1	46.2	41.3	39.1	72.3

Table 6. Ablation study of 3D-ISR module. Dist (Ort) means only using distance (orientation) to compute pairwise spatial features.

catenation (Concat.) to integrate 3D object features and 2D multi-view features along with 3D-ISR. These results indicate the synergy efforts between MCMF and 3D-ISR.

Different LLMs and Foundation Models. Table 8 presents the results of various LLMs and pre-trained 2D/3D models, including Vicuna-7B [10] vs Vicuna-13B, Ulip2 [49] vs Uni3D [59], CLIP [36] vs SigLIP [33], and ViT-H-based SAM [24] vs ViT-L-based SAM. We find that the performance of grounding and reasoning increases along with the total number of parameters in the foundation models (*i.e.* ViT-L vs Vit-H and 7B vs 13B). These results indicate that our framework’s capabilities can be improved in tandem with the performance of foundation models.

5. Conclusion and Limitations

In this paper, we proposed Inst3D-LMM, an effective instance-aware framework to leverage the potential of Large Multi-modal Models (LMMs) for 3D scene understanding. To improve instance-level representations, we developed

Method	ScanRefer		ScanQA		Scan2Cap@.50	
	Acc@0.25↑	Acc@0.50↑	B-1↑	CIDER↑	B-4↑	CIDER↑
w/ MCMF (only)	38.2	34.6	39.4	76.8	30.5	66.9
w/ 3D-ISR (only)	48.3	44.1	39.1	72.3	29.8	64.3
MCMF + PE	46.7	41.9	41.5	78.6	32.7	68.2
3D-ISR + Concat.	49.8	45.0	40.6	75.4	33.3	66.5
MCMF + 3D-ISR (full)	57.8	51.6	43.5	88.6	38.3	79.7

Table 7. Comparison results of the collaboration between MCMF and 3D-ISR modules.

LLM	3D Encoder	2D VFMs	ScanRefer		ScanQA	
			Acc@0.25↑	Acc@0.50↑	B-1↑	CIDER↑
Vicuna-7B	Ulip2	CLIP+SAM (ViT-L)	49.2	44.8	39.5	79.6
Vicuna-7B	Ulip2	CLIP+SAM (ViT-H)	53.6	46.5	43.7	87.5
Vicuna-7B	Ulip2	SigLIP+SAM (ViT-H)	54.3	47.0	42.2	84.5
Vicuna-7B	Uni3D	CLIP+SAM (ViT-H)	57.8	51.6	43.5	88.6
Vicuna-13B	Uni3D	CLIP+SAM (ViT-H)	56.0	52.0	42.8	83.1
Vicuna-13B	Uni3D	SigLIP+SAM (ViT-H)	55.2	47.8	43.3	85.4

Table 8. Ablation study of different LLMs and pre-trained foundation models.

a novel Multi-view Cross-Modal Fusion (MCMF) module, which injects multi-view 2D semantic open-vocabulary priors into 3D geometry features to generate fine-grained instance-level tokens. Furthermore, we introduced a 3D Instance Spatial Relation (3D-ISR) module that employs the spatial condition attention mechanism to capture pairwise spatial relations. Experimental results demonstrate that our approach achieves promising performance in understanding and reasoning across various 3D vision-language tasks.

Limitations. Due to the scarcity of high-quality 3D-text datasets, there remains a gap between 3D LMM learning and real-world embodied action control, such as robotic manipulation and navigation. In the future, we plan to enhance Inst3D-LMM’s reasoning and planning capabilities by scaling up diverse 3D vision and language data. Additionally, ethical safety concerns and potential hallucinatory outputs in LLM applications also warrant attention.

References

- [1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, pages 19129–19139, 2022. [5](#), [6](#), [1](#), [2](#), [3](#)
- [2] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, pages 16464–16473, 2022. [2](#)
- [3] Yang Cao, Zeng Yihan, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. In *NeurIPS*, pages 71862–71873, 2024. [4](#)
- [4] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221. Springer, 2020. [5](#), [6](#), [1](#), [2](#), [3](#)
- [5] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *NeurIPS*, pages 20522–20535, 2022. [2](#), [4](#), [5](#), [7](#)
- [6] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *CVPR*, pages 26428–26438, 2024. [1](#), [2](#), [3](#), [6](#)
- [7] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. [3](#), [6](#)
- [8] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, pages 3193–3203, 2021. [2](#), [5](#), [6](#), [1](#), [3](#)
- [9] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *ICCV*, pages 18109–18119, 2023. [2](#)
- [10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. [5](#), [8](#), [3](#)
- [11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 2432–2443, 2017. [5](#), [1](#), [2](#)
- [12] Zhipeng Ding, Xu Han, and Marc Niethammer. Votenet: A deep learning label fusion method for multi-atlas segmentation. In *MICCAI*, pages 202–210, 2019. [6](#)
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [5](#)
- [14] 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. [2](#)
- [15] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. [2](#)
- [16] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, pages 20482–20494, 2023. [1](#), [2](#), [3](#), [6](#), [7](#)
- [17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. [6](#)
- [18] 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. *arXiv preprint arXiv:2312.08168*, 2023. [3](#), [6](#), [7](#)
- [19] 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. *NeurIPS*, 2024. [1](#), [2](#), [6](#), [3](#)
- [20] 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*, 2023. [2](#), [3](#)
- [21] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, pages 4867–4876, 2020. [1](#), [2](#)
- [22] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. In *ECCV*, pages 528–545. Springer, 2022. [2](#)
- [23] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *CVPR*, pages 10984–10994, 2023. [2](#)
- [24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. [3](#), [4](#), [5](#), [8](#), [1](#)
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. [7](#), [8](#)
- [26] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024. [4](#)
- [27] Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, and Xiangde Liu. 3dmit: 3d multi-modal

- instruction tuning for scene understanding. In *ICMEW*, 2024. 6
- [28] Cai Liang, Bo Li, Zhengming Zhou, Longlong Wang, Pengfei He, Liang Hu, and Haoxing Wang. Spatio-aware-grounding3d: A spatio aware model for improving 3d vision grounding. In *CVPRW*, 2024. 4
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023. 6
- [30] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. In *NeurIPS*, pages 44860–44879, 2023. 2
- [31] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2023. 1, 2
- [32] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *arXiv preprint arXiv:2409.03757*, 2024. 2
- [33] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pages 529–544. Springer, 2022. 8
- [34] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *CVPR*, pages 5606–5611, 2023. 2
- [35] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023. 6
- [36] 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 *ICML*, pages 8748–8763, 2021. 3, 4, 5, 8, 1
- [37] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *CoRL*, pages 23–72, 2023. 1
- [38] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, pages 125–141. Springer, 2022. 5
- [39] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, pages 8216–8223, 2023. 3, 4, 5, 1, 2
- [40] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *ICCV*, pages 2998–3009, 2023. 1
- [41] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Open-Mask3D: Open-Vocabulary 3D Instance Segmentation. In *NeurIPS*, pages 68367–68390, 2023. 4
- [42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 5, 1
- [43] Ozan Unal, Christos Sakaridis, Suman Saha, Fisher Yu, and Luc Van Gool. Three ways to improve verbo-visual fusion for dense 3d visual grounding. *arXiv preprint arXiv:2309.04561*, 2023. 6
- [44] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, pages 19757–19767, 2024. 1, 3
- [45] 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. In *EMNLP*, page 10612–10625, 2023. 2
- [46] 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. 1, 3
- [47] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*, pages 19231–19242, 2023. 1
- [48] Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahu Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024. 2
- [49] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *CVPR*, pages 27091–27101, 2024. 5, 8
- [50] Fan Yang, Sicheng Zhao, Yanhao Zhang, Haoxiang Chen, Hui Chen, Wenbo Tang, Haonan Lu, Pengfei Xu, Zhenyu Yang, Jungong Han, et al. Llmi3d: Empowering llm with 3d perception from a single 2d image. *arXiv preprint arXiv:2408.07422*, 2024. 2
- [51] Jianing Yang, Xuwei 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 *ICRA*, pages 7694–7701, 2024. 6
- [52] Li Yang, Ziqi Zhang, Zhongang Qi, Yan Xu, Wei Liu, Ying Shan, Bing Li, Weiping Yang, Peng Li, Yan Wang, et al. Exploiting contextual objects and relations for 3d visual grounding. In *NeurIPS*, pages 49542–49554, 2024. 6
- [53] 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. In *NeurIPS*, pages 26650–26685, 2023. 6
- [54] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290, 2019. 6

- [55] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *CVPR*, pages 20623–20633, 2024. [6](#)
- [56] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. [3](#)
- [57] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*, pages 15225–15236, 2023. [5](#), [6](#), [1](#), [2](#), [3](#)
- [58] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvgtf: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021. [6](#)
- [59] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *ICLR*, 2024. [2](#), [3](#), [4](#), [5](#), [8](#)
- [60] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Empowering 3d visual grounding with reasoning capabilities. *arXiv preprint arXiv:2407.01525*, 2024. [1](#), [3](#), [6](#)
- [61] 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. [2](#)
- [62] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, pages 2911–2921, 2023. [2](#), [4](#), [6](#), [7](#)

Inst3D-LMM: Instance-Aware 3D Scene Understanding with Multi-modal Instruction Tuning

Supplementary Material

In this document, we provide additional experimental results and more details on our approach. The supplementary material is organized as follows:

- § 6: More implementation details;
- § 7: Additional experimental results;
- § 8: Datasets and evaluation details;
- § 9: Multi-task instruction templates.

Furthermore, we will fully release the **source code and checkpoints**. We provide example code files along with this document, including the end-to-end multi-task instruction tuning and evaluation processes for specific tasks.

6. More Implementation Details

6.1. Instance-level 3D Feature Extraction

To generate 3D instance proposals, we leverage a pre-trained transformer-based 3D instance segmentation model, Mask3D [39], which is frozen in our framework. In our method, we only utilize the predicted binary instance masks, completely ignoring the predicted class labels and confidence scores to go beyond closed-vocabulary class predictions.

6.2. Multi-view 2D Feature Extraction

To encode the text-image aligned embeddings while preserving information about long-tail or novel concepts, we utilize the CLIP [36] visual encoder to extract 2D per-instance image features. We project point clouds corresponding to each 3D instance O_i^{3D} onto the color image plane using the camera’s intrinsic (\mathbf{I}) and extrinsic ($(\mathbf{R}|t)$) matrices, resulting in 2D homogeneous coordinates $\mathbf{P}_i^{3D} = \mathbf{I} \cdot (\mathbf{R}|t) \cdot O_i^{3D}$. Considering all projected points often yield imprecise and noisy bounding boxes due to outliers, we randomly sample $k_{\text{sample}} = 5$ points and run SAM [24] for $k_{\text{rounds}} = 10$ iterations, selecting the 2D mask with the highest confidence score. This mask is then used to generate multi-level crops of the selected frame, thereby enriching the features with contextual information. For each 3D instance proposal, we collect $K \cdot L$ images by selecting the top- K views and deriving L multi-level crops.

6.3. Multi-task Instruction Tuning

We integrate text embeddings generated by a pre-trained tokenizer with instance-level attribute-aware tokens $\{O_{I1}^{3D}, O_{I2}^{3D}, \dots, O_{IN}^{3D}\}$ and global scene-level relation-aware token F_{scene} to facilitate multi-task training. For different downstream tasks, we utilize various system

messages, which are detailed in Sec. 9 of this document. The converted datasets comprise ScanRefer [4], Multi3DRefer [57], ScanQA [1], Scan2Cap [8], and the ScanNet subset of 3D-LLM [16]. The overall loss \mathcal{L}_{LLM} is formulated as follows:

$$\mathcal{L}_{LLM} = \mathcal{L}_{text} + \mathcal{L}_{SF}(O_{pred}, GT_{inst}),$$

where \mathcal{L}_{text} is the default cross-entropy loss utilized in LLaMA2 [42], while \mathcal{L}_{SF} is the sigmoid focal loss to measure the similarity between the predicted object O_{pred} and the ground-truth instance GT_{inst} . Upon completion of end-to-end multi-task instruction tuning, Inst3D-LMM can effectively handle various 3D vision-language tasks generically, without requiring specific task fine-tuning.

7. Additional Experimental Results

7.1. Ablation Studies

In addition to the ablation experiments presented in the main paper, we further provide more comprehensive results.

Effects of Input Modalities. We evaluate the influence of different input modalities as presented in Table A9. The experimental results indicate that the best performance is achieved by using both 2D images and 3D point clouds together. Relying on only one of these modalities leads to a decline in the model’s performance. Additionally, 3D point clouds are relatively more critical than 2D images.

Different 3D Instance Segmentation Models. In Table A10, we report the results obtained from employing different 3D instance segmentation models in our framework, all of which are pre-trained on the ScanNet200 [11] dataset. We compare various approaches, such as PointGroup [21] and Mask3D [39]. Furthermore, we directly incorporate the 3D ground-truth (GT) masks from annotations. One can clearly observe that our Inst3D-LMM with 3D GT masks obtains the best performance across various metrics. These results demonstrate that the performance of our Inst3D-LMM improves in conjunction with the quality of utilized 3D instance segmentation masks.

Hyperparameter K in 2D Image Views Selection. As illustrated in Table A11, we conduct an analysis of the hyperparameter K in the top- K views selection across various benchmarks. The results indicate that the performance of our approach improves as the number of 2D image frames used increases. To trade off the model’s performance with computational resource costs, we set K to 5.

Method	Inputs	ScanRefer		Multi3DRefer		ScanQA			Scan2Cap@0.50				
		Acc@0.25↑	Acc@0.50↑	F1@0.25↑	F1@0.50↑	B-1↑	METEOR↑	ROUGE↑	CIDER↑	B-4↑	METEOR↑	ROUGE↑	CIDER↑
Inst3D-LMM (2D only)	Image	33.2	28.5	31.9	26.7	29.4	10.7	27.5	57.8	18.5	16.9	39.3	58.6
Inst3D-LMM (3D only)	Point Cloud	48.3	44.1	46.2	41.3	39.1	14.6	34.4	72.3	29.8	23.4	50.4	64.3
Inst3D-LMM (full)	Image&Point Cloud	57.8	51.6	58.3	53.5	43.5	18.4	42.6	88.6	38.3	27.5	57.2	79.7

Table A9. Ablation results with different input modalities.

Pre-trained 3D Model	ScanRefer		Multi3DRefer		ScanQA			Scan2Cap@0.50				
	Acc@0.25↑	Acc@0.50↑	F1@0.25↑	F1@0.50↑	B-1↑	METEOR↑	ROUGE↑	CIDER↑	B-4↑	METEOR↑	ROUGE↑	CIDER↑
PointGroup [21]	53.2	47.5	53.2	49.6	39.4	17.6	39.7	83.5	35.6	26.6	56.1	75.8
Mask3D [39]	57.8	51.6	58.3	53.5	43.5	18.4	42.6	88.6	38.3	27.5	57.2	79.7
3D GT Masks	65.2	58.7	66.9	58.5	49.8	25.5	48.6	92.3	44.7	33.6	63.5	86.3

Table A10. Ablation study with different pre-trained 3D instance segmentation models.

Top- K	ScanRefer		Multi3DRefer		ScanQA			Scan2Cap@0.50				
	Acc@0.25↑	Acc@0.50↑	F1@0.25↑	F1@0.50↑	B-1↑	METEOR↑	ROUGE↑	CIDER↑	B-4↑	METEOR↑	ROUGE↑	CIDER↑
1	52.4	46.8	48.2	44.3	32.6	10.5	30.4	77.6	28.5	22.4	46.9	68.8
5	57.8	51.6	58.3	53.5	43.5	18.4	42.6	88.6	38.3	27.5	57.2	79.7
10	57.6	52.2	57.6	54.3	42.8	18.6	43.1	86.3	38.4	27.0	55.8	79.2

Table A11. Ablation study on the hyperparameter K of the top- K views selection.

7.2. Main Comparisons

Specialist Model vs. Generalist Model. To assess the effectiveness of our generalist approach in distinguishing between different tasks using multi-modal instructions, we evaluate our model across various tasks. Table A12 reports the experimental results. The first four rows present the performance of models trained from scratch as specialized experts for each dataset. The final row indicates our model trained with a generalist approach. The results demonstrate that our model can effectively distinguish between multiple tasks such as 3D-VG, 3D-QA, and 3D-DC, outperforming the specialist models in most benchmarks without specific task training.

Training&Inference Efficiency. To demonstrate the efficiency of our method, we provide detailed comparisons between our Inst3D-LMM and other 3D LMMs, such as LEO [20], LL3DA [6] and Chat-Scene [19]. As shown in Table A13, our Inst3D-LMM requires fewer computational resources while offering faster training and inference speeds. Additionally, we evaluate the training time and memory usage with and without the integration of multi-view 2D information. The results indicate that including multi-view 2D priors slightly increases training time by 0.5 hours and VRAM usage per GPU by 5 GB.

7.3. More Visualization Results

We present additional visualization results on various tasks, such as 3D Visual Grounding (Figure A8), 3D Question Answering (Figure A9), 3D Dense Captioning (Figure A10) and 3D Scene Descriptions (Figure A11). We also extend our model to potential robot applications by leveraging embodied dialogue and planning datasets from 3D-

LLM [16], as illustrated in the visualizations for Embodied Dialogue (Figure A12) and Embodied Planning (Figure A13). These qualitative results demonstrate the capability of Inst3D-LMM to enhance robot manipulation and navigation in real-world applications.

8. Datasets and Evaluation Details

ScanRefer [4]: This benchmark evaluates the performance of 3D object localization through natural language descriptions from the ScanNetv2 dataset [11]. The scenes are labeled as “unique” if they only contain one object of a class, or “multiple” if they contain more. The official assessment of prediction accuracy employs the percentage of predictions achieving IoU above 0.25 and 0.5 with the corresponding ground-truth bounding boxes. Inst3D-LMM selects the object with the highest similarity score between the predicted object O_{pred} and the ground-truth instance GT_{inst} to facilitate fair comparisons against previous works.

Multi3DRefer [57]: This dataset focuses on localizing multiple objects in real-world 3D scenes using natural language descriptions from the ScanNetv2 dataset. Performance evaluation hinges on calculating the F1 score at IoU thresholds of 0.25 and 0.5. The official metric adopts the Hungarian algorithm to optimally match the predicted bounding boxes with their ground-truth.

ScanQA [1]: This dataset is derived from the ScanNetv2 dataset, which is annotated for the 3D Question Answering task to assess visual and spatial understanding of 3D environments. It enables models to answer text-based queries about 3D scans. We utilize *BLEU-1*, *METEOR*, *ROUGE*, and *CIDER* metrics as evaluation protocols.

Scan2Cap [8]: The dense captioning benchmark, based on

Method	ScanRefer		Multi3DRefer		ScanQA				Scan2Cap@0.50			
	Acc@0.25↑	Acc@0.50↑	F1@0.25↑	F1@0.50↑	B-4↑	METEOR↑	ROUGE↑	CIDER↑	B-4↑	METEOR↑	ROUGE↑	CIDER↑
ScanRefer [4]	56.9	51.4	—	—	—	—	—	—	—	—	—	—
Multi3DRefer[57]	—	—	57.8	53.8	—	—	—	—	—	—	—	—
ScanQA [1]	—	—	—	—	14.5	18.8	41.9	87.2	—	—	—	—
Scan2Cap [8]	—	—	—	—	—	—	—	—	38.6	26.5	57.8	78.6
Inst3D-LMM	57.8	51.6	58.3	53.5	14.9	18.4	42.6	88.6	38.3	27.5	57.2	79.7

Table A12. Comprehensive evaluations as a generalist method. The first four rows present the performance of models trained from scratch as specialized experts for each dataset. The final row evaluates our model trained with a generalist approach. As a generalist, our Inst3D-LMM excels in distinguishing each task and achieving superior results across most metrics.

Method	LLM	VRAM per GPU↓	Training Time↓	Inference Time↓
LEO [20]	Vicuna-7B [10]	~80 GB	~42 hours	~4.80 seconds
LL3DA [6]	OPT-6.7B [56]	~60 GB	~36 hours	~1.20 seconds
Chat-Scene [19]	Vicuna-7B	~70 GB	~8 hours	~0.96 seconds
Inst3D-LMM (3D only)	Vicuna-7B	~50 GB	~3.5 hours	~0.48 seconds
Inst3D-LMM (3D+2D)	Vicuna-7B	~55 GB	~4 hours	~0.52 seconds

Table A13. Comparison of memory usage, training and inference time between our Inst3D-LMM and other methods. Here, we define the average model forward time for a single instruction on the ScanQA validation dataset as the inference time.

ScanRefer, requires models to detect objects and generate captions simultaneously. The linguistic generation is evaluated using *BLEU-4*, *METEOR*, *ROUGE*, and *CIDER* metrics, weighted by IoU scores above 0.25 or 0.5 with ground-truth bounding boxes.

ScanNet subset of 3D-LLM. [16]: In addition to the object-centric dataset, we utilize the ScanNet subset of 3D-LLM for scene-level 3D scene descriptions. This task requires the model to translate its comprehensive understanding of the 3D scene into natural languages. We have enhanced our Inst3D-LMM to freely reference objects with identifiers when describing a complex 3D scene.

9. Multi-task Instruction Templates

Inst3D-LMM simultaneously handles various vision tasks, such as visual grounding, question answering, dense captioning, and scene descriptions. Each task requires distinct and diverse templates to enable the LLM to provide accurate responses, tailored to the specific requirements of each benchmark. We provide several example instruction templates for each task.

3D Visual Grounding. We present our instruction tuning templates for single object grounding on the ScanRefer dataset, as shown in Figure A14, and for multiple objects grounding on the Multi3DRefer dataset, as depicted in Figure A15. Additionally, we also consider the scenario where no object corresponds to the given query, which is indicated as “No object”.

3D Question Answering. As illustrated in Figure A16, for the question-answering task on the ScanQA dataset, we append suffixes to the model’s outputs to indicate whether they consist solely of phrases or single words, in accordance with the annotation guidelines of the dataset.

3D Dense Captioning. As depicted in Figure A17, the question templates in the Scan2Cap dataset require the model to describe the visual attributes of the targeted object while also exploring its spatial relationships with other elements in the scene.

3D Scene Descriptions. We utilize question templates, as shown in Figure A18, to enable our Inst3D-LMM to effectively translate its comprehensive understanding of the entire 3D scene into natural language on the ScanNet subset of 3D-LLM [16].

Embodied Dialogue&Planning. We further extend the model to include embodied dialogue and embodied planning. As illustrated in Figure A19 and Figure A20, the templates prompt the model to generate concise, natural responses or clear, step-by-step plans based on given instruction information.



Figure A8. More qualitative results for 3D Visual Grounding on ScanRefer validation dataset. The rendered images of 3D scans are presented, including the ground-truth (blue), incorrectly identified objects (red), and correctly identified objects (green). The colored text indicates the results of text decoupling. Please zoom in for better visualization (the same as below).

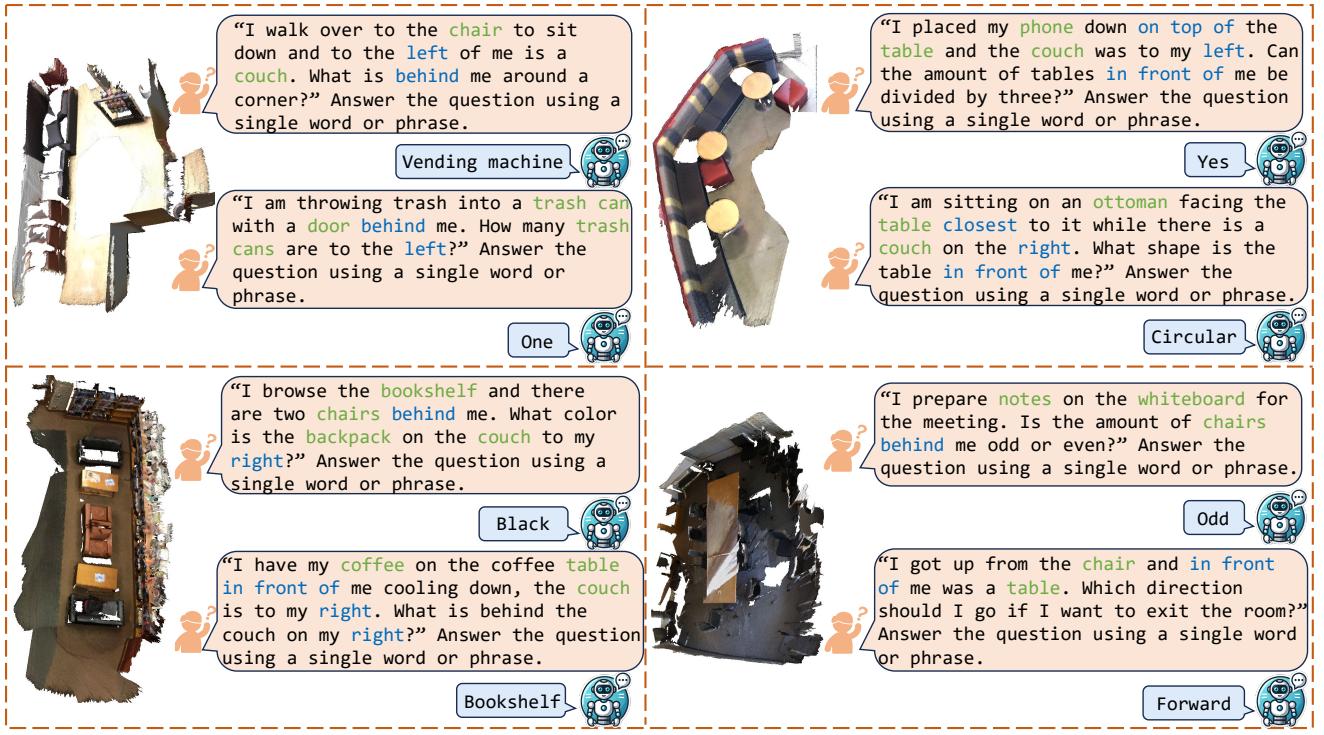


Figure A9. More qualitative results for 3D Question Answering on ScanQA validation dataset. The colored text indicates the results of text decoupling.

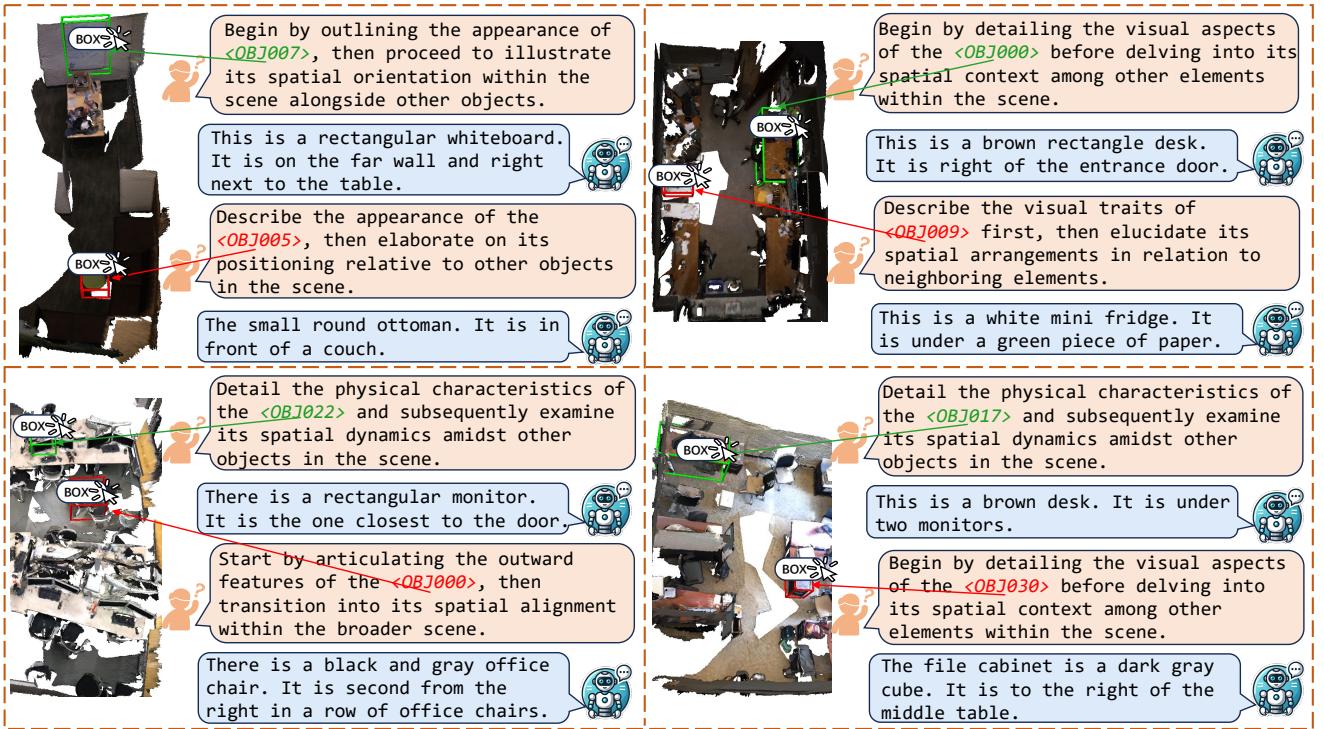


Figure A10. More qualitative results for 3D Dense Captioning on Scan2Cap validation dataset (with grounding).

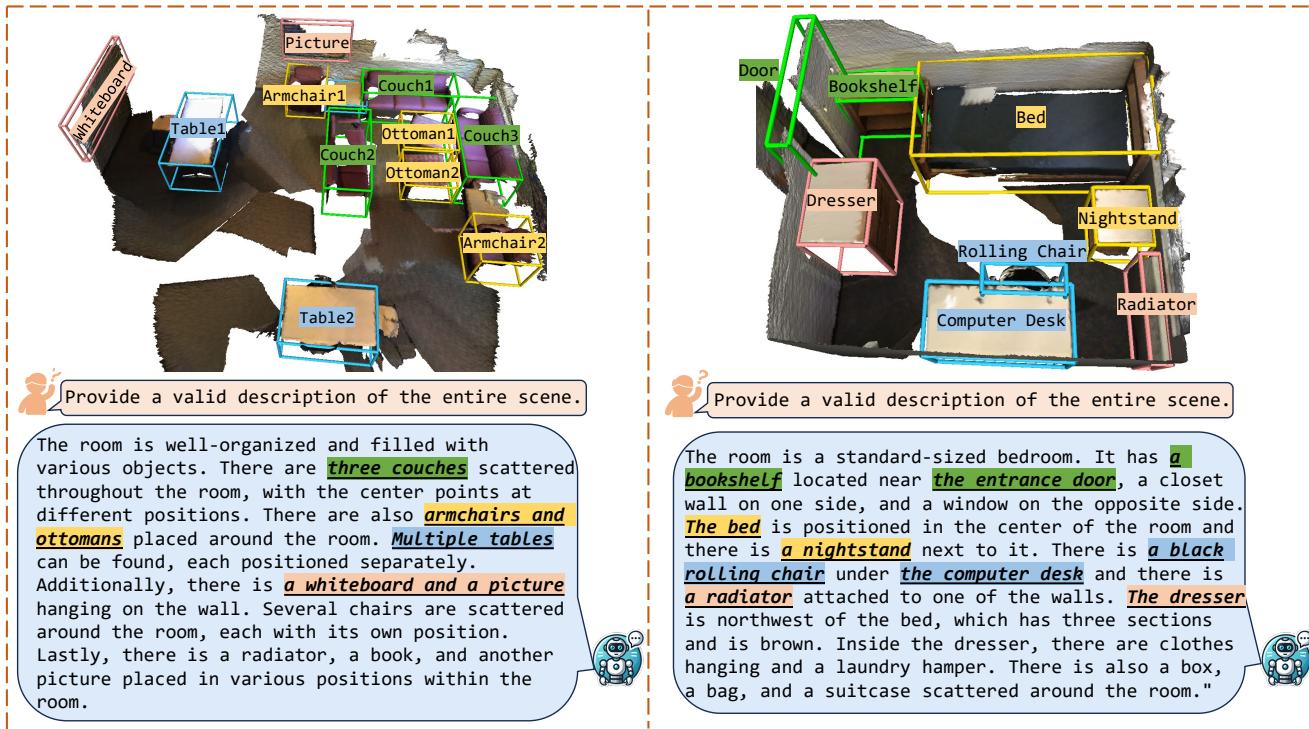


Figure A11. More qualitative results for 3D Scene Descriptions on ScanNet subset of 3D-LLM.

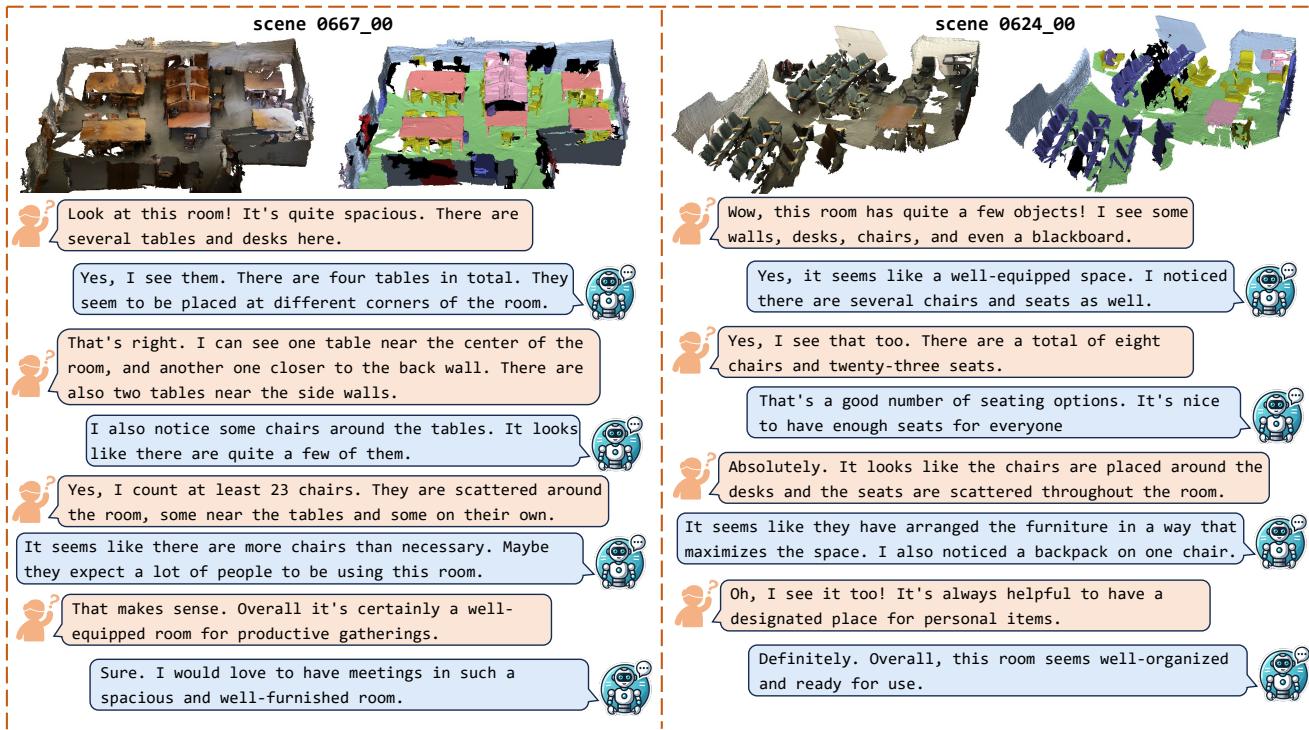


Figure A12. More qualitative results for Embodied Dialogue. Please zoom in for better visualization.

scene 0015_00



I want to set up a meeting room. what should I do?

1. place chairs around the table in a semi-circle.
2. place a podium in front of the chairs.
3. hang a whiteboard on the wall.
4. position the table in the center of the room.
5. place a cabinet against the wall.
6. arrange the remaining chairs along the walls.
7. install windows on the designated wall locations

I want to set up a meeting room. I have done these things:
 1. place chairs around the table
 2. position the cabinet against the wall for storage
 3. put the whiteboard on the wall for presentations.
 4. arrange the podium near the whiteboard for speakers
 what should I do next?

open the windows for fresh air and natural light

scene 0164_00



I want to clean the kitchen. what should I do?

1. open the kitchen cabinet
2. take out the cleaning supplies
3. close the kitchen cabinet
4. clean the kitchen counter
5. clean the stove
6. clean the kitchen sink
7. clean the refrigerator
8. put the cleaning supplies back in the kitchen cabinet
9. close the kitchen cabinet

I want to prepare a meal in the kitchen. I have done these things:
 1. open the refrigerator
 2. take out the necessary ingredients.
 what should I do next?

go to the kitchen counter

scene 0195_00



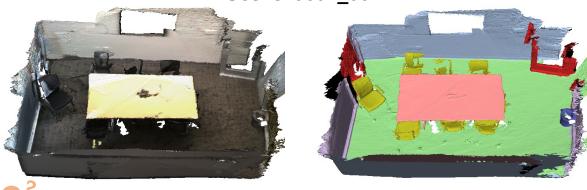
I want to clean the bathroom. what should I do?

1. go to the bathroom cabinet
2. take a towel from the cabinet
3. go to the sink
4. turn on the water
5. wash hands and face
6. turn off the water
7. use the soap dispenser to wash hands
8. use the towel to dry hands and face
9. go to the toilet
10. use the toilet paper
11. throw away the used toilet paper
12. leave the bathroom

I want to clean up the bathroom. I have done these things:
 1. gather towels from the towel rack
 2. clean the sink using the soap dispenser and water.
 what should I do next?

clean the mirror using a towel

scene 0601_00



I want to clean up the office space. what should I do?

1. gather trash from the desk and put it in the trash can
2. collect recyclable items and place them in the bin
3. organize and tidy up items on the desk
4. clear any clutter from the office chairs
5. wipe down the desk and chairs to clean them
6. empty the trash can into the larger trash can or dumpster
7. dust off any surfaces, including the dresser and shelves
8. clean the telephone with a disinfectant wipe
9. open the window to let in fresh air
10. close the window after cleaning is complete

I want to clean up the office. I have done these things:
 1. pick up trash and place it in the trash can
 2. collect recyclables and place them in the bin
 3. organize items on the desk.
 what should I do next?

arrange books and items on the shelf

Figure A13. More qualitative results for Embodied Planning. Please zoom in for better visualization.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Please share the ID of the object that closely matches the description <description>.

<Inst3D-LMM>: Here is the <OBJ_ID> based on your request.

Question templates:

- "Please share the ID of the object that closely matches the description <description>."
- "Provide the ID of the object that best fits the description <description>."
- "What is the ID of the object that aligns with the description <description>?"
- "Identify the ID of the object that closely resembles the description <description>."
- "What's the ID of the object that corresponds to the description <description>?"
- "Give the ID of the object that most accurately matches the description <description>."
- "Share the ID of the object that best corresponds to the description <description>."
- "Identify the ID of the object that closely aligns with the description <description>."

Answer templates:

Single object:

- "Here is the <OBJ_ID> based on your request."
- "The grounding process has identified this <OBJ_ID>."
- "I've successfully pinpointed the <OBJ_ID>."

No object:

- "There is no object based on your description."
- "No object detected in this indoor scene."
- "No object found in the indoor setting."

Figure A14. Instruction templates for 3D Visual Grounding (single object grounding).

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details.

Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>].

Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Are there any objects fitting the description <description>? If so, please provide their IDs.

<Inst3D-LMM>: Yes, there are <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> within the interior.

Question templates:

- "Are there any objects fitting the description <description>? If so, please provide their IDs."
- "Do any objects match the description <description>? If they do, share their IDs."
- "Is there anything that matches the description <description>? If yes, provide their IDs."
- "Are there objects that correspond to the description <description>? If so, kindly list their IDs."
- "Does anything fit the description <description>? If it does, list the IDs of those objects."
- "Have any objects been described as <description>? If so, share their IDs."
- "Do any objects meet the criteria of <description>? If they do, kindly provide their IDs."
- "Are there any objects that correspond to the description <description>? If yes, share their IDs."

Answer templates:

Multiple objects:

- "Yes, there are <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> within the interior."
- "Detected <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> in this indoor scene."
- "Several <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> have been identified in this indoor setting."

No object:

- "There are no objects based on your description."
- "No objects detected in this indoor scene."
- "No objects found in the indoor setting."

Figure A15. Instruction templates for 3D Visual Grounding (multiple objects grounding).

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: "What is under the sink?" Please answer the question using a single word or phrase.

<Inst3D-LMM>: Kitchen cabinets. The ID of the object is <OBJ015>.

Question templates:

- "<Raw Question> Please answer the question using a single word or phrase."

Answer templates:

- "<Raw Answer> The ID of the object is <OBJ_ID>."

Figure A16. Instruction templates for 3D Question Answering.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Start by detailing the visual aspects of the <OBJ_ID>, then delve into its spatial context within the scene.

<Inst3D-LMM>: <Raw Answer>

Question templates:

- "Start by detailing the visual aspects of the <OBJ_ID>, then delve into its spatial context within the scene."
- "Outline the appearance of the <OBJ_ID>, then elaborate on its positioning relative to other objects in the scene."
- "Illustrate the visual attributes of the <OBJ_ID>, then explore its spatial relationships with other elements in the scene."
- "Begin by articulating the outward features of the <OBJ_ID>, then discuss its spatial alignment within the broader scene."
- "Provide a detailed description of the <OBJ_ID>'s appearance before analyzing its spatial connections with other elements in the scene."
- "Capture the essence of the <OBJ_ID>'s appearance, then analyze its spatial relationships within the scene's context."
- "Detail the physical characteristics of the <OBJ_ID>, then examine its spatial dynamics among other objects in the scene."

Answer templates:

- "<Raw Answer>"

Figure A17. Instruction templates for 3D Dense Captioning.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Provide a valid description of the entire scene.

<Inst3D-LMM>: The 3D scene depicts a versatile living space with various objects, thoughtfully designed to blend functionality with aesthetic appeal. In one corner, a black microwave (OBJ001) sits above brown cabinets (OBJ023)...

Question templates:

- "Provide a valid description of the entire scene."

Answer templates:

- "<Raw Answer>"

Figure A18. Instruction templates for 3D Scene Descriptions.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: "Look at this room! It's quite spacious. There are several tables and desks here." The answer should be concise and natural. Do not add any objects or details not mentioned in the instruction message.

<Inst3D-LMM>: Yes, I see them. There are four tables in total. They seem to be placed at different corners of the room.

Question templates:

- "<Raw Question> The answer should be concise and natural. Do not add any objects or details not mentioned in the instruction information."

Answer templates:

- "<Raw Answer>"

Figure A19. Instruction templates for Embodied Dialogue.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: "I want to clean the bathroom. what should I do?" Please give a clear, concise step-by-step plan. Each step should be simple and under 15 words. Do not add any objects or details not mentioned in the instruction message.

<Inst3D-LMM>: 1.go to the bathroom cabinet 2.take a towel from the cabinet 3.go to the sink 4.turn on the water 5.wash hands and face 6.turn off the water 7. use the soap dispenser to wash hands...

Question templates:

- "<Raw Question> Please give a clear, concise step-by-step plan. Each step should be simple and under 15 words. Do not add any objects or details not mentioned in the instruction information."

Answer templates:

- "<Raw Answer>"

Figure A20. Instruction templates for Embodied Planning.