PAR3D: A Unified 3D-MLLM with Part-Aware Representation for Scene Understanding

Shaohui Dai* **Yansong Qu*†** **You Shen** **Shengchuan Zhang** **Liujuan Cao‡**
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University

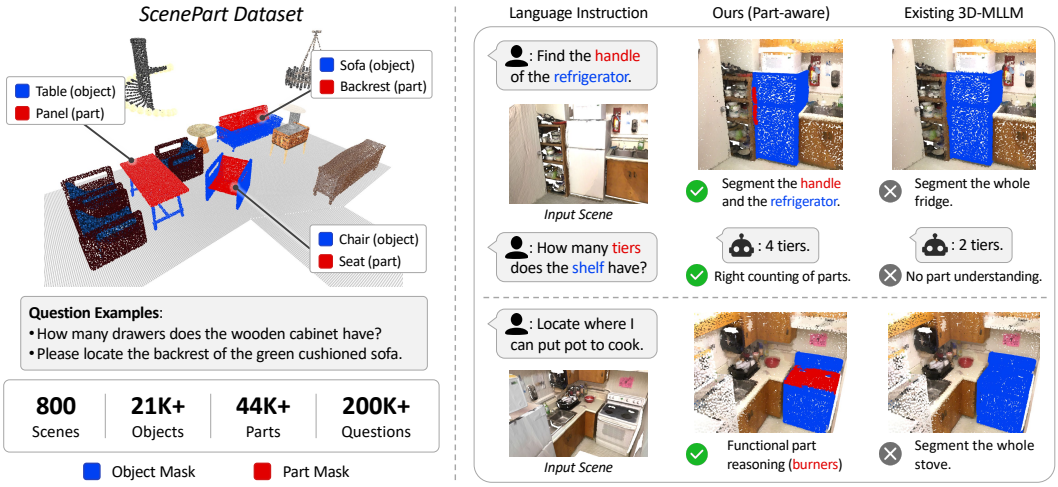

Figure 1: We propose PAR3D, a unified 3D-MLLM with part-aware representation, together with ScenePart dataset. *Left:* ScenePart provides fine-grained object-part annotations and language instructions for 3D scenes. *Right:* PAR3D enables part-aware understanding across question answering, segmentation, and reasoning, going beyond the object-level understanding of existing 3D-MLLMs.

Abstract

Recent advances in 3D multimodal large language models (3D-MLLMs) have enabled unified solutions for 3D scene understanding tasks, including visual question answering, captioning, and referring segmentation. However, existing 3D-MLLMs remain largely object-centric, limiting their ability to model fine-grained part structures that are essential for embodied interaction with 3D environments. In this work, we present *PAR3D*, a unified part-aware 3D-MLLM framework that enables models to understand, reason about, and ground both objects and their parts in 3D scenes. To enable training and evaluation of part-aware 3D scene understanding, we introduce ScenePart, a synthetic 3D scene dataset with part-level annotations and language instructions. We further develop Part-Aware 3D Representation Learning to enrich 3D visual representations with fine-grained part-level semantics, and propose Hierarchical Segmentation Query Generation to ground part targets via hierarchical object–part queries. Extensive experiments show that our method substantially improves part-level question answering and referring segmentation,

*Equal contribution.
†Project Lead.
‡Corresponding Author

Preprint.

while also achieving strong performance across object-level vision-language tasks. Project page: https://atrovast.github.io/PAR3D/.

1 Introduction

Language-guided understanding of 3D environments is a fundamental problem in computer vision and a key capability for embodied intelligence. Recent 3D multimodal large language models (3D-MLLMs) have made substantial progress toward this goal by connecting 3D perception modules with large language models, enabling a unified interface for 3D scene perception and spatial reasoning [35, 64, 24, 23]. These advances make 3D-MLLMs promising for robotics, augmented reality, and digital twins, where intelligent systems must interpret complex scenes and respond to language instructions.

However, many real-world interactions require understanding a scene beyond object-level recognition. An embodied agent may need to grasp the *handle* of a mug, pull the *drawer* of a cabinet, sit on the *seat* of a chair, or inspect the *door* of a refrigerator, requiring affordance-aware perception and manipulation of object parts [38, 19]. Beyond embodied intelligence, similar fine-grained structural awareness is also useful for controllable 3D content editing and interactive scene manipulation, where systems may need to select or modify local functional regions rather than entire objects [46, 52]. In these cases, the target of reasoning is not merely an object instance, but a functional component embedded within an object and situated in a scene. This requires part-aware 3D scene understanding: the model should recognize object parts as meaningful semantic and functional units, understand their dependence on host objects and scene context, and localize target parts within host objects.

Part-aware 3D scene understanding exposes a structural mismatch in existing 3D-MLLMs. Most current approaches are built around an object-centric view of 3D scenes, where objects serve as the primary units for visual representation [23, 10, 63], language alignment [21, 24, 18], and task supervision [15, 20, 26]. A straightforward solution is to extend the object-centric formulation to object parts as finer-grained targets. Object parts, however, are not simply smaller objects. They are structured components whose meanings are tied to their host objects and whose functions often determine how an agent should interact with them. As illustrated in Figure 1, existing 3D-MLLMs may identify the host object but overlook its functional parts. This object-centric bias manifests in three key aspects. First, existing 3D datasets [13, 51, 5] largely lack part-level annotations in scene context, making it difficult to supervise object–part understanding. Second, existing 3D visual backbones are typically adapted from object-level scene perception and may not preserve the fine-grained geometric and semantic cues needed to distinguish object parts [24, 15]. Third, grounding mechanisms often rely on a single-granularity query formulation, forcing object-level and part-level targets to share the same query representation and weakening part-aware grounding [15, 26, 72].

To address these challenges, we propose PAR3D, a part-aware 3D-MLLM framework for unified object- and part-level understanding in 3D scenes. Instead of treating parts as merely finer-grained object targets, our framework models them as functional components embedded within objects and situated in scene context. Concretely, we improve part-aware 3D understanding from three aspects: data, representation, and grounding. First, we construct **ScenePart**, a synthetic 3D scene dataset that places part-annotated objects into realistic indoor scenes and provides object and part masks, object–part correspondences, and language instructions in scene context. Second, we develop **Part-Aware 3D Representation Learning** on top of a pretrained 3D foundation encoder, adapting the visual backbone to capture fine-grained geometric and semantic features of object parts. Third, we introduce **Hierarchical Segmentation Query Generation**, which generates decoupled segmentation queries for object- and part-level targets to support grounding of parts in relation to their host objects. Together, these designs form a part-aware 3D-MLLM framework that supports diverse 3D vision-language tasks over both objects and their parts.

In summary, our main contributions are as follows:

- We construct ScenePart, a synthetic scene-level dataset with object and part masks, object–part correspondences, and language annotations for part-aware 3D-MLLM training and evaluation.
- We propose a part-aware 3D-MLLM framework with Part-Aware 3D Representation Learning and Hierarchical Segmentation Query Generation, enabling fine-grained understanding in 3D scenes.
- Extensive experiments demonstrate that our framework improves both spatial understanding and grounding at object and part levels across multiple 3D vision-language tasks.

2 Related Works

2.1 3D Vision-Language Models for 3D Scenes

3D vision-language models connect 3D environments with natural language for tasks such as grounding, description, and question answering. Early studies mainly formulate these tasks with dedicated benchmarks and specialist models. For example, ScanRefer [7] and ReferIt3D [1] study language-guided object grounding in 3D scenes, Scan2Cap [11] addresses dense caption generation, and ScanQA [2] and SQA3D [34] introduce question answering over 3D environments. Beyond these task-specific formulations, subsequent specialist models further improve 3D-language modeling through representation pretraining, stronger cross-modal fusion, and joint modeling across related tasks [73, 12, 8, 56]. In parallel, recent efforts on semantic-aware neural scene representations, including NeRF- and 3DGS-based methods, also explore how to embed semantic, language, or open-vocabulary information into continuous 3D representations [43, 45, 14, 27, 44, 54, 55]. These methods establish important foundations for 3D language understanding.

With the development of large language models, recent 3D-MLLMs extend language-grounded 3D perception toward instruction following and open-ended reasoning. These methods connect LLMs with 3D scenes through rendered multi-view images, object token representations, or direct point-cloud features [21, 72, 23, 60, 48, 42, 25, 69]. Recent unified frameworks further support multiple 3D vision-language tasks within a single model [9, 71, 15, 26]. Despite their growing generality, these methods still largely rely on object-level supervision and use objects as the primary units for visual representation, language alignment, and grounding. As a result, they lack explicit modeling of object–part hierarchy and remain limited in fine-grained part-aware understanding in 3D scenes.

2.2 3D Part Perception

3D part perception aims to identify semantically meaningful components of individual 3D objects. Early works mainly study supervised part segmentation on such objects, using part-annotated datasets such as ShapeNetPart [6, 62] and PartNet [37]. Classic supervised methods [41, 28, 40, 68, 57] learn point-level representations to predict predefined part categories on single-object point clouds. Recent studies improve scalable 3D part segmentation by lifting 2D foundation-model predictions to 3D [70, 61], distilling 2D vision-language priors into 3D representations [49], and constructing large-scale 3D part supervision [36, 33]. These efforts enable promptable, open-set, or text-aligned part decomposition for 3D objects. Interaction-oriented studies in embodied AI [59, 38, 19, 53] further explore object parts for affordance reasoning, articulation modeling, and manipulation.

Despite these advances, existing approaches mainly focus on part segmentation or decomposition of individual 3D objects, leaving part-aware understanding in complete 3D scenes underexplored. In contrast, our work studies object parts in complete 3D scenes, where parts need to be understood in relation to their host objects and scene context. We incorporate part-level supervision into a unified 3D-MLLM for fine-grained reasoning and grounding.

3 Method

We present PAR3D, a part-aware unified 3D-MLLM framework for multiple 3D vision-language tasks over objects and parts. Given a colored point cloud $\mathbf{X} \in \mathbb{R}^{N\times 6}$ and a language instruction, the model generates either a textual response or segmentation masks for referred objects or parts. Our framework supports part-aware 3D understanding through three components: the proposed dataset **ScenePart** provides scene-level part supervision, **Part-Aware 3D Representation Learning** adapts 3D visual backbone to better capture fine-grained features, and **Hierarchical Segmentation Query Generation** produces granularity-aware grounding tokens for object- and part-level mask prediction.

3.1 Preliminary: 3D-LLaVA

Our framework builds on 3D-LLaVA [15], a 3D-MLLM that unifies language-based 3D scene understanding and referring segmentation. Given a point cloud $\mathbf{X}$, a 3D encoder $\mathcal{E}$ extracts point-level features and aggregates them into superpoint-level features:

$$\mathbf{F}_e = \mathrm{Pool}(\mathcal{E}(\mathbf{X})), \tag{1}$$

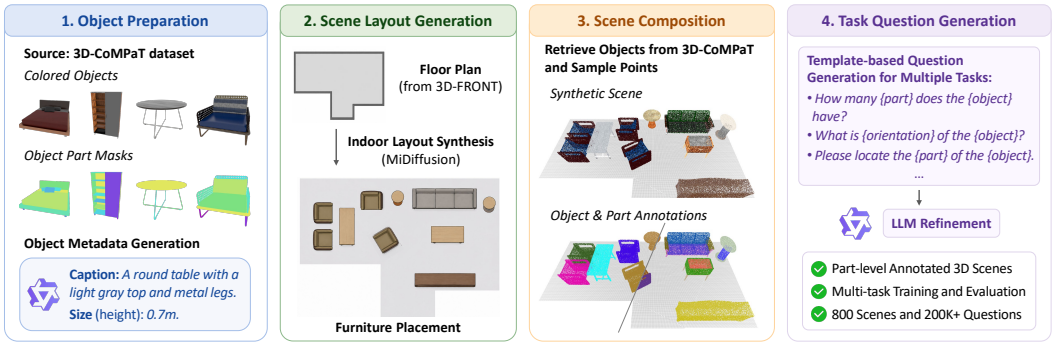

Figure 2: **ScenePart Data Construction Pipeline.** ScenePart composes part-annotated 3D objects into synthesized indoor layouts, producing object- and part-level mask annotations in 3D scenes and multi-task language instructions for training and evaluating part-aware 3D-MLLMs.

where $\mathrm{Pool}(\cdot)$ denotes superpoint pooling, and $\mathbf{F}_e$ consists of M encoder features $\{\mathbf{f}_i^e\}_{i=1}^M$ over superpoints. A query decoder $\mathcal{D}$ further refines the superpoint features, using $\mathbf{F}_e$ as both the visual input queries and their key-value features:

$$\mathbf{F}_d = \mathcal{D}(\mathbf{F}_e, \mathbf{F}_e), \tag{2}$$

where $\mathbf{F}_d$ consists of refined features $\{\mathbf{f}_i^d\}_{i=1}^M$ over the same superpoints. An MLP projector $\mathcal{P}(\cdot)$ maps the decoder features into the LLM embedding space:

$$\mathbf{V} = \mathcal{P}(\mathbf{F}_d), \tag{3}$$

where $\mathbf{V}$ denotes the 3D visual tokens provided to the LLM.

To enable referring segmentation, 3D-LLaVA uses a special token $[\mathrm{SEG}]$. Its hidden state serves as a mask query and is decoded with the 3D visual features to predict a target mask over superpoints. While effective for object-level 3D scene understanding, this framework remains limited in capturing fine-grained part-aware semantics. We extend this framework toward part-aware 3D scene understanding by improving its visual representation and segmentation query design.

3.2 ScenePart Dataset

Scene-level part supervision is largely missing in existing 3D vision-language data. Indoor scene datasets provide realistic object-level annotations [13, 51, 5] and language tasks [7, 2, 11, 34, 65], while part-annotated datasets mainly focus on isolated objects [37, 32, 29, 47]. To support part-aware 3D-MLLM training, we construct **ScenePart**, a synthetic 3D scene dataset that places part-annotated objects into indoor scenes and provides object- and part-level supervision in scene context.

Each ScenePart scene is represented as a colored point cloud with object masks, part masks, and object–part correspondences that associate each part instance with its host object. We also provide object descriptions and construct a scene graph that records spatial relationships among objects. In total, ScenePart contains 800 scenes with 21K object masks, 44K part masks, and 273K language-task annotations. Based on these annotations, we build **ScenePart-200K** as the training set, which contains both visual question answering and referring segmentation instructions. For evaluation, we further construct two test splits: **ScenePart-QA** evaluates part-aware 3D question answering with diverse question types, while **ScenePart-Seg** evaluates referring segmentation across objects and parts at different granularities. In our framework, ScenePart supports both representation adaptation with dense object and part masks, and instruction tuning with language-task annotations.

We construct ScenePart in four steps, as illustrated in Figure 2. First, we preprocess part-annotated 3D assets from 3D-CoMPaT [29, 47] by filtering object models and normalizing their sizes. We use Qwen3-VL-8B [3] to estimate object scales and generate object descriptions that are later used for language annotation. Second, we generate indoor layouts using MiDiffusion [22] on floor plans from 3D-FRONT [16], obtaining furniture placements with category, position, orientation, and scale. Third, we instantiate these placements with the preprocessed assets and sample them into point-cloud scenes, where object masks cover the inserted object instances, and part masks are inherited from their

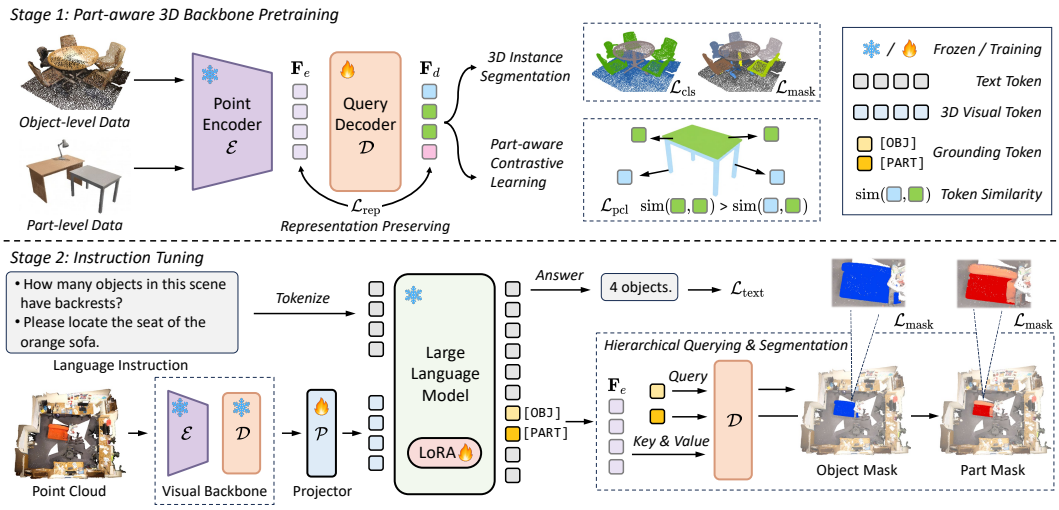

Figure 3: **Overall Framework of PAR3D.** PAR3D is trained with a two-stage scheme. Stage 1 adapts the 3D visual backbone with object- and part-level supervision through instance segmentation, part-aware contrastive learning, and representation-preserving regularization. Stage 2 performs instruction tuning on the MLLM using 3D vision-language instruction data. PAR3D generates textual responses and object or part masks through hierarchical grounding tokens [OBJ] and [PART].

internal part annotations. Finally, we generate language-task annotations from object descriptions, part semantics, object–part correspondences, and scene-graph relations. We use template-based rules for controllability and LLM-based refinement for linguistic diversity.

3.3 Part-Aware 3D Representation Learning

Part-aware 3D understanding requires a visual backbone that can capture both scene-level semantics and fine-grained part geometry. Existing 3D-MLLMs commonly rely on 3D encoders or visual backbones adapted from object-level scene understanding, which provide strong object-level perception but offer limited support for fine-grained part representation. To obtain a more general 3D representation, we instantiate the 3D encoder $\mathcal{E}$ with a pretrained Point Transformer model [57, 58, 66, 67]. This encoder provides rich geometric and semantic priors, offering a strong foundation for fine-grained recognition in 3D scenes.

In our unified 3D-MLLM, the visual backbone consists of a frozen pretrained encoder $\mathcal{E}$ and a trainable query decoder $\mathcal{D}$. The decoder produces features $\mathbf{F}_d$, which are used for constructing visual tokens. A strong encoder alone does not guarantee effective representations for the LLM, as the decoder adapts the features to the model's downstream tasks. During visual backbone training, we observe that the adapted decoder features can become biased toward mask prediction, potentially deviating from the semantic structure encoded by the frozen encoder. To address this, we introduce two regularization objectives for visual backbone training: Part-aware contrastive learning leverages ScenePart annotations to enhance part-level separability in $\mathbf{F}_d$, while representation-preserving self-distillation aligns the adapted decoder features with the pretrained encoder features.

Part-aware contrastive learning. Given a ScenePart scene, we apply contrastive learning to the decoder features $\mathbf{F}_d$ at the superpoint level, compacting features within the same part while separating features from different parts. Let $\mathcal{S}_k$ denote the superpoint indices covered by the k-th part mask. For an anchor feature $\mathbf{f}_i^d$, $\mathcal{P}(i)$ contains decoder features of other superpoints in the same part mask, and $\mathcal{N}(i)$ contains decoder features of superpoints in different part masks. The part-aware contrastive loss follows an InfoNCE objective:

$$\mathcal{L}_{\text{pcl}} = -\sum_i \log \frac{\sum_{j\in\mathcal{P}(i)} \exp(\text{sim}(\mathbf{f}_i^d, \mathbf{f}_j^d)/\tau)}{\sum_{j\in\mathcal{P}(i)\cup\mathcal{N}(i)} \exp(\text{sim}(\mathbf{f}_i^d, \mathbf{f}_j^d)/\tau)}, \tag{4}$$

where τ is the temperature and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. This objective improves intra-part feature consistency and strengthens the separability of different part regions.

Representation-preserving self-distillation. While part-aware contrastive learning introduces fine-grained supervision, visual backbone adaptation may shift $\mathbf{F}_d$ toward task-specific segmentation representations and weaken the general 3D semantics provided by the pretrained encoder. To control this shift, we use the frozen encoder features $\mathbf{F}_e$ as semantic anchors for the decoder features $\mathbf{F}_d$.

Specifically, for each superpoint, we encourage the decoder feature to remain close to its corresponding encoder feature in the normalized feature space:

$$\mathcal{L}_{\mathrm{rep}} = 1 - \frac{1}{M}\sum_{i=1}^{M} \mathrm{sim}\left(\mathbf{f}_i^d, \mathrm{sg}(\mathbf{f}_i^e)\right), \tag{5}$$

where M is the number of superpoints and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. This objective regularizes the visual backbone adaptation by preserving the semantic structure of the pretrained encoder and reducing the task-specific drift of $\mathbf{F}_d$.

3.4 Hierarchical Segmentation Query Generation

Part-aware grounding requires the language model to generate grounding tokens for targets at different granularities, ranging from whole objects to object parts. However, existing 3D-MLLMs with referring segmentation commonly use a single grounding token $[\mathrm{SEG}]$ to represent referred targets of all granularities [15, 26]. As a result, object-level and part-level targets are represented by the same type of query, which can cause granularity conflicts and weaken fine-grained part grounding. Moreover, part-level grounding requires object-part context, as a part is defined with respect to its host object rather than as an independent region.

We introduce Hierarchical Segmentation Query Generation to reduce this granularity conflict and expose object–part structure in the language-to-grounding interface. Instead of using one generic grounding token for all targets, the LLM generates granularity-aware grounding tokens. For object-level grounding, it generates an object token $[\mathrm{OBJ}]$ to represent the referred object. For part-level grounding, it generates $[\mathrm{OBJ}]$ followed by $[\mathrm{PART}]$, so that host object and target part are represented by separate grounding tokens while remaining coupled in the same language context.

Let $\mathbf{h}_{\mathrm{obj}}$ and $\mathbf{h}_{\mathrm{part}}$ denote the hidden states of $[\mathrm{OBJ}]$ and $[\mathrm{PART}]$, respectively. We project these hidden states into segmentation queries with a lightweight MLP $\phi(\cdot)$:

$$\mathbf{s}_{\mathrm{obj}} = \phi(\mathbf{h}_{\mathrm{obj}}), \qquad \mathbf{s}_{\mathrm{part}} = \phi(\mathbf{h}_{\mathrm{part}}), \tag{6}$$

We reuse the same query decoder $\mathcal{D}$ for mask decoding. Specifically, the derived segmentation queries are decoded with the encoder features $\mathbf{F}_e$ as key-value features. The decoder predicts superpoint-level masks, which are further mapped back to point-level masks:

$$\hat{\mathbf{m}}_{\mathrm{obj}} = \mathcal{D}(\mathbf{s}_{\mathrm{obj}}, \mathbf{F}_e), \qquad \hat{\mathbf{m}}_{\mathrm{part}} = \mathcal{D}(\mathbf{s}_{\mathrm{part}}, \mathbf{F}_e). \tag{7}$$

During instruction tuning, we supervise the response format according to the target granularity. Object-level referring expressions are trained to generate $[\mathrm{OBJ}]$ and predict the corresponding object mask. Part-level referring expressions are trained to generate both $[\mathrm{OBJ}]$ and $[\mathrm{PART}]$, with mask supervision on the host object and the target part. This joint token and mask supervision encourages the model to preserve object–part structure in its generated grounding tokens and derived segmentation queries while maintaining a unified grounding interface for both object-level and part-level tasks.

3.5 Training Strategy

We train the model in two stages, as illustrated in Figure 3. The first stage performs part-aware 3D backbone pretraining, where the query decoder is adapted with object- and part-level supervision while the pretrained point encoder remains frozen. The second stage conducts instruction tuning, where the visual backbone is frozen and the projector and LLM are optimized to support multiple 3D vision-language tasks over objects and their parts.

Table 1: **Quantitative Comparison on Object-Level Benchmarks.** We compare state-of-the-art methods across 3D referring segmentation, question answering, and dense captioning. The best and second-best results are highlighted in **bold** and underlined, respectively.

Methods	ScanRefer (val)	Multi3DRefer (val)	ScanQA (val)				SQA3D (test)		Scan2Cap (val)			
	mIoU↑	mIoU↑	C↑	B-4↑	M↑	R-L↑	EM↑	EM-R↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R-L@0.5↑
Specialist Models:												
ScanQA[2]	-	-	64.9	10.1	13.1	33.3	46.6	-	-	-	-	-
3D-VisTA[73]	-	-	69.6	10.4	13.9	**45.7**	48.5	-	61.6	34.1	26.8	55.0
Scan2Cap[11]	-	-	-	-	-	-	41.0	-	39.1	23.3	22.0	44.8
UniT3D[12]	-	-	-	-	-	-	-	-	46.7	27.2	21.9	46.0
Vote2Cap-DETR [8]	-	-	-	-	-	-	-	-	61.8	34.5	26.2	54.4
M3DRef-CLIP [65]	35.7	32.6	-	-	-	-	-	-	-	-	-	-
3D-STMN [56]	39.5	-	-	-	-	-	-	-	-	-	-	-
Finetuned 3D-MLLMs:												
3D-LLM [21]	-	-	69.4	12.0	14.5	35.7	-	-	-	-	-	-
Scene-LLM [18]	-	-	80.0	12.0	16.8	40.0	54.2	-	-	-	-	-
LL3DA [9]	-	-	76.8	13.5	15.9	37.3	-	-	65.2	36.8	26.0	55.1
SegPoint [20]	41.7	36.1	-	-	-	-	-	-	-	-	-	-
Generalist 3D-MLLMs:												
LEO [24]	-	-	101.4	13.2	20.0	49.2	50.0	52.4	72.4	**38.2**	27.9	**58.1**
Scene-LLM [18]	-	-	80.0	11.7	15.8	35.9	53.6	-	-	-	-	-
Reason3D [26]	42.0	-	73.5	12.1	15.1	37.4	-	-	-	-	-	-
Chat-Scene [23]	-	-	87.7	14.3	18.0	41.6	54.6	**57.5**	77.2	36.4	**28.0**	**58.1**
Grounded 3D-LLM [10]	-	-	72.7	13.4	-	-	-	-	70.6	35.5	-	-
3DGraphLLM [63]	-	-	88.8	15.9	-	-	**55.9**	-	81.0	36.5	-	-
3D-LLaVA [15]	43.3	42.7	92.6	**17.1**	18.4	43.1	54.5	56.6	78.8	36.9	27.1	57.7
PAR3D (ours)	**49.9**	**53.4**	**95.7**	15.9	**18.9**	45.0	54.6	57.3	**81.4**	37.3	27.5	57.9

Stage 1: part-aware 3D backbone pretraining. In the first stage, we train the query decoder $\mathcal{D}$ while keeping the pretrained point encoder $\mathcal{E}$ frozen. This stage uses two types of 3D supervision from ScanNet and ScenePart. On ScanNet, we train the query decoder with the 3D instance segmentation objective, which maintains its object-level mask prediction ability:

$$\mathcal{L}_{\text{inst}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mask}}. \tag{8}$$

Here, $\mathcal{L}_{\text{cls}}$ supervises instance category prediction, and $\mathcal{L}_{\text{mask}}$ supervises the predicted instance masks. On ScenePart, we use the part-level annotations to apply the part-aware contrastive loss $\mathcal{L}_{\text{pcl}}$ introduced in Section 3.3. In addition, for all training scenes in this stage, we apply the representation-preserving loss $\mathcal{L}_{\text{rep}}$. The overall objective for stage 1 is:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{inst}} + \lambda_{\text{pcl}}\mathcal{L}_{\text{pcl}} + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}}, \tag{9}$$

where λ_{pcl} and λ_{rep} balance the auxiliary objectives.

Stage 2: instruction tuning. In the second stage, we freeze the 3D visual backbone and train the projector $\mathcal{P}$ and the segmentation MLP ϕ together with the LLM using LoRA. We combine existing 3D language understanding datasets with our ScenePart-200K to construct the instruction-tuning data. The existing datasets include ScanRefer [7], Nr3D [1], Multi3DRefer [65], ScanQA [2], SQA3D [34], and Scan2Cap [11], covering object-level referring segmentation, question answering, and captioning. ScenePart further introduces part-aware question answering and referring segmentation instructions.

For language-only tasks, we optimize the standard autoregressive text loss $\mathcal{L}_{\text{text}}$. For grounding tasks, the model is supervised to generate the corresponding grounding tokens and predict masks. Object-level referring expressions generate $[\text{OBJ}]$ and are supervised with object mask annotations, while part-level referring expressions generate both $[\text{OBJ}]$ and $[\text{PART}]$ and are supervised with the host object mask and the target-part mask. We denote the corresponding grounding supervision as $\mathcal{L}_{\text{mask}}$. The stage 2 objective is:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{text}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{10}$$

where the mask loss term is omitted for tasks without grounding supervision.

4 Experiments

4.1 Implementation Details

The LLM backbone of our work is LLaVA-1.5-7B [31], following the setting of 3D-LLaVA [15]. Specifically, the original 3D encoder is replaced with the pretrained Utonia encoder [67]. The training proceeds in two stages. In Stage 1, the visual backbone is trained on ScanNet200 [13] and ScenePart

Table 2: **Quantitative Comparison on the ScenePart Benchmark.** We compare PAR3D with representative 3D-MLLMs on ScenePart-Seg and ScenePart-QA, covering referring segmentation at different granularities and visual question answering. The best results are highlighted in **bold**.

Methods	ScenePart-Seg								ScenePart-QA			
	Object		Coarse-Part		Fine-Part		All		C↑	B-4↑	M↑	R-L↑
	mIoU↑	Acc@0.5↑	mIoU↑	Acc@0.5↑	mIoU↑	Acc@0.5↑	mIoU↑	Acc@0.5↑				
3D-LLaVA [15]	29.0	23.9	9.0	3.4	4.2	1.0	11.1	6.6	39.6	0.1	10.0	32.1
Reason3D [26]	25.5	13.8	7.4	2.3	3.9	0.8	9.6	4.0	-	-	-	-
3D-LLaVA +ScenePart	78.4	83.3	52.8	55.7	37.6	37.4	51.8	53.9	177.2	45.6	43.7	80.6
PAR3D (ours)	**89.6**	**91.5**	**60.9**	**66.0**	**46.0**	**47.1**	**60.7**	**63.5**	**191.1**	**61.7**	**46.9**	**82.1**

for 256 epochs using AdamW with an initial learning rate of 3×10^{-4} and polynomial decay. In Stage 2, we perform LoRA-based instruction tuning for 2 epochs on the instruction datasets described in Sec. 3.5, keeping the main LLM parameters frozen and optimizing the projector MLP and LoRA parameters with AdamW, using an initial learning rate of 2×10^{-4} and a cosine annealing schedule. All experiments are conducted on 4 NVIDIA A100 GPUs with 40GB memory.

4.2 Datasets and Metrics

Datasets. We evaluate our method on the proposed ScenePart test splits and several 3D vision-language benchmarks. For part-aware evaluation, we use ScenePart-Seg and ScenePart-QA introduced in Section 3.2. ScenePart-Seg evaluates referring segmentation over objects and parts at different granularities. ScenePart-QA evaluates part-aware 3D question answering with diverse question types involving part semantics, part counting, spatial relations, and object–part associations. To examine object-level performance, we also report results on ScanRefer [7] and Multi3DRefer [65] for object-level referring segmentation, ScanQA [2] and SQA3D [34] for 3D question answering, and Scan2Cap [11] for dense caption generation. These benchmarks are based on ScanNet scenes and have been widely adopted in previous works.

Metrics. For referring segmentation tasks, we report mean Intersection-over-Union (mIoU) between predicted and ground-truth masks. On ScenePart-Seg, we further report Acc@0.5, i.e., the percentage of predictions with IoU above 50%, and present results separately for object, coarse-part, and fine-part targets. For open-ended language-generation tasks, including question answering and captioning, we report CIDEr, BLEU-4, METEOR, and ROUGE-L following standard evaluation protocols, abbreviated as C, B-4, M, and R-L in the tables. For SQA3D, which follows a definite-answer setting, we report exact-match accuracy (EM) and refined exact-match accuracy (EM-R) following common evaluation practice. Higher values indicate better performance for all metrics.

4.3 Comparison with State-of-the-Art Methods

We compare PAR3D with existing state-of-the-art approaches across multiple 3D vision-language tasks, covering both object-level and part-level evaluation. The compared methods are grouped into three categories. *Specialist models* are task-specific methods designed for a particular benchmark or a narrow set of closely related tasks. Most of them do not rely on LLMs. *Finetuned 3D-MLLMs* are adapted to each dataset or task through task-specific finetuning and often achieve strong in-domain performance. *Generalist 3D-MLLMs* are trained once on a diverse collection of 3D vision-language tasks and then evaluated across multiple datasets without task-specific finetuning. PAR3D falls into the generalist category and further extends generalist 3D-MLLMs with part-awareness.

Object-level evaluation. Table 1 compares different methods on object-level benchmarks, including 3D referring segmentation, question answering, and dense captioning. Overall, PAR3D achieves strong performance across the three groups of compared methods while operating as a generalist 3D-MLLM. For 3D referring segmentation, PAR3D achieves new state-of-the-art results among 3D-MLLMs on both ScanRefer and Multi3DRefer, outperforming the previous best generalist model, 3D-LLaVA, by absolute gains of 6.6% and 10.7%, respectively. For language-generation tasks, PAR3D also shows strong performance across question answering and dense captioning. On SQA3D, ScanQA, and Scan2Cap, it achieves competitive performance across most evaluated metrics among all compared methods. LEO [24]'s ScanQA results are marked in gray and excluded from the main comparison, as they are obtained under a different setting with access to question-related ground-

truth objects. These results demonstrate that the proposed framework improves object-level scene understanding as well, further strengthening the unified multi-task capability of 3D-MLLMs.

Part-level evaluation. Since existing benchmarks do not evaluate part-aware 3D scene understanding, we conduct part-aware evaluation on the proposed ScenePart test datasets, as shown in Table 2. We compare PAR3D with representative 3D-MLLMs that support referring segmentation, including 3D-LLaVA and Reason3D. Reason3D is evaluated only on ScenePart-Seg, as its released model focuses on segmentation and does not support open-ended textual answering for QA tasks. We also train 3D-LLaVA with ScenePart-200K as a stronger baseline to separate the effect of ScenePart supervision from our model design. For ScenePart-Seg, we report mIoU and Acc@0.5 at multiple granularities, including object-level targets, coarse parts (*e.g.*, door of a refrigerator), fine-grained parts (*e.g.*, handle of a refrigerator), together with the overall result. For ScenePart-QA, we report standard language-generation metrics. Across these settings, PAR3D consistently outperforms representative 3D-MLLMs, demonstrating stronger part-aware 3D scene understanding. Notably, training 3D-LLaVA with ScenePart-200K improves its part-aware performance, indicating that ScenePart provides effective supervision. Nevertheless, PAR3D further improves over this data-enhanced baseline, showing the benefit of the proposed part-aware model design.

Qualitative results. Figure 4 presents representative examples on real ScanNet scenes with part-aware instructions for referring segmentation and visual question answering. Although part-level supervision is only provided by our synthetic ScenePart dataset during training, PAR3D can generalize to real-world 3D scans. In segmentation, blue and red masks indicate object- and part-level predictions, respectively. PAR3D predicts fine-grained part masks with host-object context, while 3D-LLaVA often produces coarser object masks or misses the referred part. In part-aware question answering, PAR3D correctly answers a part-aware counting question by recognizing object components such as pillow and bucket handle. These examples qualitatively demonstrate PAR3D's ability to handle fine-grained object-part understanding across referring segmentation and question answering. Additional qualitative comparisons are provided in Appendix D.

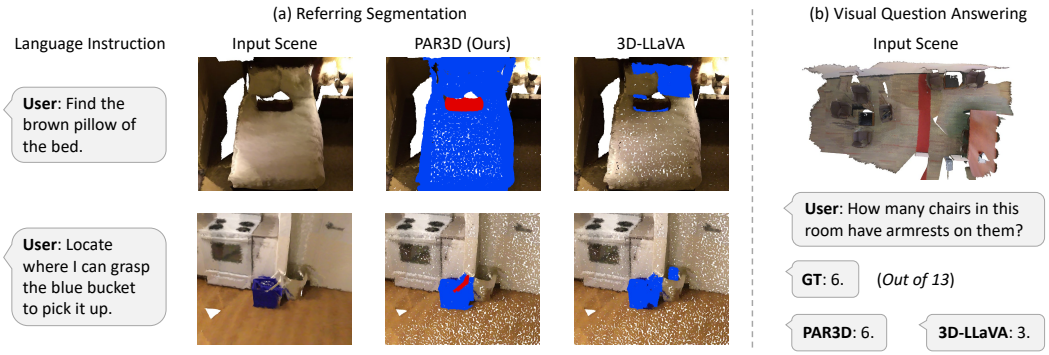

Figure 4: **Qualitative Examples of PAR3D.** We present examples on (a) referring segmentation and (b) visual question answering for part-aware 3D scene understanding.

4.4 Ablation Studies

We conduct ablation studies to analyze the contribution of each component in our framework. Since our main designs focus on visual representation learning and hierarchical segmentation query generation, we report mIoU on ScenePart-Seg and ScanRefer to evaluate part- and object-level grounding. As shown in Table 3, training 3D-LLaVA with ScenePart-200K improves ScenePart-Seg mIoU from 11.1% to 51.8%, highlighting the importance of ScenePart supervision for part-aware learning. The improvement on ScanRefer indicates that incorporating ScenePart does not harm object-level grounding. Beyond data supervision, part-aware 3D representation learning further improves both benchmarks. The pretrained 3D encoder provides a stronger visual foundation, while the representation-preserving loss and part-aware contrastive loss further adapt the features for fine-grained representations. Finally, hierarchical segmentation query generation achieves the best results on both ScenePart-Seg and ScanRefer, demonstrating that using separate yet coupled segmentation queries for objects and parts benefits fine-grained part grounding as well as object-level grounding.

Table 3: **Ablation Studies on Referring Segmentation.** We evaluate the contribution of each component on ScenePart-Seg and ScanRefer, covering part-aware and object-level referring segmentation. The best results are highlighted in **bold**.

Methods	ScenePart-Seg mIoU↑	ScanRefer mIoU↑
3D-LLaVA (baseline)	11.1	43.3
+ ScenePart Data	51.8	44.4
+ Pretrained 3D Encoder	54.9	47.4
+ Representation-Preserving Loss	58.7	49.1
+ Part-Aware Contrastive Loss	59.4	49.2
+ Hierarchical Segmentation Query	**60.7**	**49.9**

5 Conclusion

We introduce PAR3D, a unified 3D-MLLM with part-aware representation for fine-grained 3D scene understanding. To extend 3D understanding beyond object level, we construct ScenePart, a synthetic scene-level dataset with object and part masks, object–part correspondences, and language instructions. Building on ScenePart, we develop part-aware 3D representation learning to improve fine-grained part representations, and introduce hierarchical segmentation query generation to ground part targets via hierarchical queries. Experiments on ScenePart and standard 3D vision-language benchmarks show that PAR3D substantially improves part-aware question answering and segmentation while maintaining strong object-level performance. We hope this work encourages future unified 3D-MLLMs toward fine-grained perception, reasoning, and grounding over objects and their parts.

Limitations Although PAR3D improves part-aware 3D scene understanding, several limitations remain. ScenePart provides object- and part-level supervision through synthesized indoor scenes, but may still have a domain gap from real-world 3D scans and is limited by the source object and part categories. Future work may extend PAR3D to real-scene annotations, open-vocabulary part categories, and embodied scenarios requiring more complex object-part reasoning and interaction.

References

[1] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer, 2020.

[2] D. Azuma, T. Miyanishi, S. Kurita, and M. 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.

[3] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] S. Banerjee and A. 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.

[5] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

[6] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[7] D. Z. Chen, A. X. Chang, and M. Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[8] S. Chen, H. Zhu, X. Chen, Y. Lei, G. Yu, and T. 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.

[9] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. 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.

[10] Y. Chen, S. Yang, H. Huang, T. Wang, R. Xu, R. Lyu, D. Lin, and J. Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024.

[11] Z. Chen, A. Gholami, M. Nießner, and A. 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.

[12] Z. Chen, R. Hu, X. Chen, M. Nießner, and A. X. Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18109–18119, 2023.

[13] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. 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.

[14] S. Dai, Y. Qu, Z. Li, X. Li, S. Zhang, and L. Cao. Training-free hierarchical scene understanding for gaussian splatting with superpoint graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3673–3682, 2025.

[15] J. Deng, T. He, L. Jiang, T. Wang, F. Dayoub, and I. Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3772–3782, 2025.

[16] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021.

[17] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, and D. Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129(12):3313–3337, 2021.

[18] R. Fu, J. Liu, X. Chen, Y. Nie, and W. Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

[19] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7081–7091, 2023.

[20] S. He, H. Ding, X. Jiang, and B. Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pages 349–367. Springer, 2024.

[21] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

[22] S. Hu, D. M. Arroyo, S. Debats, F. Manhardt, L. Carlone, and F. Tombari. Mixed diffusion for 3d indoor scene synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1262–1272, 2026.

[23] H. Huang, Y. Chen, Z. Wang, R. Huang, R. Xu, T. Wang, L. Liu, X. Cheng, Y. Zhao, J. Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

[24] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

[25] J. Huang, X. Ma, X. Linghu, Y. Fan, J. He, W. Tan, Q. Li, S.-C. Zhu, Y. Chen, B. Jia, et al. Leo-vl: Efficient scene representation for scalable 3d vision-language learning. *arXiv preprint arXiv:2506.09935*, 2025.

[26] K.-C. Huang, X. Li, L. Qi, S. Yan, and M.-H. Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In *2025 International Conference on 3D Vision (3DV)*, pages 1177–1186. IEEE, 2025.

[27] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. LERF: language embedded radiance fields. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19672–19682. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01807. URL `https://doi.org/10.1109/ICCV51070.2023.01807`.

[28] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

[29] Y. Li, U. Upadhyay, H. Slim, A. Abdelreheem, A. Prajapati, S. Pothigara, P. Wonka, and M. Elhoseiny. 3d compat: Composition of materials on parts of 3d things. In *European conference on computer vision*, pages 110–127. Springer, 2022.

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

[31] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[32] M. Liu, Y. Zhu, H. Cai, S. Han, Z. Ling, F. Porikli, and H. Su. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21736–21746, 2023.

[33] C. Ma, Y. Li, X. Yan, J. Xu, Y. Yang, C. Wang, Z. Zhao, Y. Guo, Z. Chen, and C. Guo. P3-sam: Native 3d part segmentation. *arXiv preprint arXiv:2509.06784*, 2025.

[34] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[35] X. Ma, B. Smart, Y. Bhalgat, S. Chen, X. Li, J. Ding, J. Gu, D. Z. Chen, S. Peng, J.-W. 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.

[36] Z. Ma, Y. Yue, and G. Gkioxari. Find any part in 3d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7818–7827, 2025.

[37] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019.

[38] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[39] K. Papineni, S. Roukos, T. Ward, and W.-J. 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.

[40] A. V. Phan, M. Le Nguyen, Y. L. H. Nguyen, and L. T. Bui. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, 108:533–543, 2018.

[41] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[42] Z. Qi, Y. Fang, Z. Sun, X. Wu, T. Wu, J. Wang, D. Lin, and H. 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.

[43] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister. Langsplat: 3d language gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 20051–20060. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01895. URL `https://doi.org/10.1109/CVPR52733.2024.01895`.

[44] Y. Qu, Y. Wang, and Y. Qi. Sg-nerf: Semantic-guided point-based neural radiance fields. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 570–575. IEEE, 2023.

[45] Y. Qu, S. Dai, X. Li, J. Lin, L. Cao, S. Zhang, and R. Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5328–5337, 2024.

[46] Y. Qu, D. Chen, X. Li, X. Li, S. Zhang, L. Cao, and R. Ji. Drag your gaussian: Effective drag-based editing with score distillation for 3d gaussian splatting. *ArXiv preprint*, abs/2501.18672, 2025. URL `https://arxiv.org/abs/2501.18672`.

[47] H. Slim, X. Li, Y. Li, M. Ahmed, M. Ayman, U. Upadhyay, A. Abdelreheem, A. Prajapati, S. Pothigara, P. Wonka, et al. 3dcompat++: An improved large-scale 3d vision dataset for compositional recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[48] Y. Tang, X. Han, X. Li, Q. Yu, Y. Hao, L. Hu, and M. 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.

[49] A. Umam, C.-K. Yang, M.-H. Chen, J.-H. Chuang, and Y.-Y. Lin. Partdistill: 3d shape part segmentation by vision-language model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2024.

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

[51] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019.

[52] C. Wang, J. Ye, Y. Yang, Y. Li, Z. Lin, J. Zhu, Z. Chen, Y. Luo, and C. Guo. Part-x-mllm: Part-aware 3d multimodal large language model. *arXiv preprint arXiv:2511.13647*, 2025.

[53] J. Wang, D. Wang, J. Hu, Q. Zhang, J. Yu, and L. Xu. Kinematify: Open-vocabulary synthesis of high-dof articulated objects. *arXiv preprint arXiv:2511.01294*, 2025.

[54] Y. Wang, J. Wang, Y. Qu, and Y. Qi. Rip-nerf: learning rotation-invariant point-based neural radiance field for fine-grained editing and compositing. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 125–134, 2023.

[55] Y. Wang, J. Wang, R. Gao, Y. Qu, W. Duan, S. Yang, and Y. Qi. Look at the sky: Sky-aware efficient 3d gaussian splatting in the wild. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[56] C. Wu, Y. Ma, Q. Chen, H. Wang, G. Luo, J. Ji, and X. Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5940–5948, 2024.

[57] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024.

[58] X. Wu, D. DeTone, D. Frost, T. Shen, C. Xie, N. Yang, J. Engel, R. Newcombe, H. Zhao, and J. Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22193–22204, 2025.

[59] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[60] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024.

[61] Y. Yang, Y. Huang, Y.-C. Guo, L. Lu, X. Wu, E. Y. Lam, Y.-P. Cao, and X. Liu. Sampart3d: Segment any part in 3d objects. *arXiv preprint arXiv:2411.07184*, 2024.

[62] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

[63] T. Zemskova and D. Yudin. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8885–8895, 2025.

[64] J. Zha, Y. Fan, X. Yang, C. Gao, and X. Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025.

[65] Y. Zhang, Z. Gong, and A. 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.

[66] Y. Zhang, X. Wu, Y. Lao, C. Wang, Z. Tian, N. Wang, and H. Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. *arXiv preprint arXiv:2510.23607*, 2025.

[67] Y. Zhang, X. Wu, Y. Yang, X. Fan, H. Li, Y. Zhang, Z. Huang, N. Wang, and H. Zhao. Utonia: Toward one encoder for all point clouds. *arXiv preprint arXiv:2603.03283*, 2026.

[68] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[69] D. Zheng, S. Huang, and L. Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8995–9006, 2025.

[70] Y. Zhou, J. Gu, T. Y. Chiang, F. Xiang, and H. Su. Point-sam: Promptable 3d segmentation model for point clouds. *arXiv preprint arXiv:2406.17741*, 2024.

[71] C. Zhu, T. Wang, W. Zhang, K. Chen, and X. Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision*, pages 151–168. Springer, 2024.

[72] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4295–4305, 2025.

[73] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. 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.

A Evaluation Metrics

We evaluate PAR3D with existing approaches on referring segmentation, visual question answering, and dense captioning tasks. Since these tasks involve different output formats, we adopt task-specific evaluation metrics following prior 3D scene-language benchmarks.

Referring segmentation metrics. For referring segmentation, including ScenePart-Seg, ScanRefer [7], and Multi3DRefer [65], we evaluate the overlap between the predicted mask and the ground-truth mask using intersection-over-union (IoU). For ScenePart-Seg, we additionally report Acc@0.5, which measures the percentage of samples whose predicted mask achieves an IoU greater than 50% with the ground-truth mask.

Question answering metrics. For visual question answering, we follow the common evaluation protocols of the benchmarks in the 3D-MLLM literature. For ScenePart-QA and ScanQA [2], we report CIDEr [50], METEOR [4], ROUGE-L [30], and BLEU-4 [39]. CIDEr measures consensus between generated answers and reference answers using weighted n-gram similarity. METEOR evaluates text similarity through word alignments based on exact, stem, synonym, and paraphrase matches. ROUGE-L measures similarity based on the longest common subsequence while BLEU-4 evaluates modified n-gram precision up to 4-grams. For SQA3D [34], we report EM and EM-R [24]. EM measures the percentage of predictions that exactly match the ground-truth answers, while EM-R follows a refined exact-match protocol that performs more flexible answer matching.

Dense captioning metrics. For dense captioning on Scan2Cap [11], we report CIDEr, METEOR, ROUGE-L, and BLEU-4 with the IoU threshold of 0.5, denoted as CIDEr@0.5, METEOR@0.5, ROUGE-L@0.5, and BLEU-4@0.5. The suffix @0.5 indicates that the captioning score is counted only when the predicted object region matches the ground-truth object with IoU above 50%. These metrics jointly evaluate whether the model can both localize the target object and generate a caption that matches the reference descriptions.

B Details on ScenePart Dataset Construction

ScenePart is constructed to provide scene-level supervision for object parts, which is largely absent from existing 3D scene-language datasets. It integrates part-annotated objects into synthesized indoor scenes and provides object masks, part masks, object–part correspondences, scene graphs, and language-task annotations. The construction of ScenePart proceeds in four steps, following the pipeline illustrated in Fig. 2.

First, we preprocess part-annotated 3D assets from 3D-CoMPaT [29, 47]. We filter object models according to category compatibility and annotation quality, normalize their geometry, and preserve their part labels and object–part correspondences. We also use Qwen3-VL-8B [3] to estimate object scales, which are used to match and instantiate assets for the room layouts, and to generate object descriptions, which support downstream language-task annotation.

Second, we generate indoor layouts using MiDiffusion [22], a diffusion-based scene synthesis model trained on 3D-FRONT [16]. Given a floor plan, MiDiffusion predicts furniture placements, where each placement specifies the object category, spatial position, orientation, and scale. We use floor plans from 3D-FRONT to obtain diverse scene-level spatial configurations.

Third, we instantiate the generated layouts with the preprocessed 3D assets and sample them into complete point-cloud scenes. For each furniture placement, we retrieve a compatible object model according to its category and geometry. For categories where part-annotated assets are unavailable, we use 3D-FUTURE [17] models to preserve scene completeness. During layout instantiation, we further perform category-aware augmentation by manually inserting and placing additional compatible objects to enrich scene diversity and part-category coverage. These objects are selected from the preprocessed asset pool according to room type, object scale and spatial compatibility, and are placed in plausible locations under collision constraints. Object-level masks are inherited from individual object instances, while part-level masks are preserved only from 3D-CoMPaT models with part annotations and maintained together with their corresponding host object masks. We further construct a scene graph from the spatial relationships among objects, which provides structured scene context for annotation generation.

Finally, we generate language-task annotations from the synthesized scenes. For referring segmentation, we create expressions that refer to either whole objects or object parts and provide the corresponding object or part masks as supervision. For visual question answering, we generate questions based on object descriptions, part semantics, object–part correspondences, and spatial relationships derived from the scene graph. We first use template-based rules to ensure correctness and controllability, and then apply LLM-based refinement to improve linguistic diversity and naturalness.

Overall, this process yields complete scene-level annotations that jointly support object-level and part-level supervision. In addition to object masks and part masks, ScenePart explicitly records the correspondence between each part mask and its host object mask. This paired object–part supervision enables models to learn hierarchical grounding, where a target part is grounded together with the object to which it belongs. These annotations are then used to construct the training set and the held-out evaluation splits described in the following section.

C Statistics and Splits of ScenePart

Dataset scale. ScenePart contains 800 synthesized indoor scenes, 21K object masks, 44K part masks, and 273K language-task annotations. These annotations cover both object-level and part-level reasoning and grounding tasks. From the generated annotation pool, we sample 200K annotations for training and construct two held-out evaluation splits with 10K samples each.

Table 4: **Overall Statistics of ScenePart.**

Statistic	Count
Scenes	800
Object masks	21K
Part masks	44K
Language-task annotations	273K
Splits:	
ScenePart-200K annotations	200K
ScenePart-QA samples	10K
ScenePart-Seg samples	10K

Table 5: **Breakdown of the ScenePart-Seg and ScenePart-QA Test Splits.**

ScenePart-Seg		ScenePart-QA	
Type	Count	Type	Count
Object	2K	Part existence	3K
Coarse-part	4K	Part counting	1.5K
Fine-part	4K	Part color	2K
		Part spatial relation	2K
		Cross-object	1.5K

Train/test split. We split ScenePart at the scene level to avoid leakage between training and evaluation. ScenePart-200K is constructed from annotations associated with the training scenes. The held-out scenes are used to build two evaluation splits: ScenePart-QA for part-aware visual question answering and ScenePart-Seg for referring segmentation at multiple granularities.

ScenePart-Seg. ScenePart-Seg evaluates language-guided mask prediction at three granularities: object, coarse part, and fine part. Object-level expressions refer to complete object instances. Coarse and fine parts are defined according to the part hierarchy provided by 3D-CoMPaT, where coarse parts correspond to higher-level functional components (*e.g.*, door of a refrigerator) and fine parts correspond to lower-level subparts (*e.g.*, handle of a refrigerator). Each sample consists of a scene, a referring expression, and the ground-truth mask of the target object or part.

ScenePart-QA. ScenePart-QA evaluates part-aware 3D visual question answering. Its questions are generated from object names, part labels, object–part correspondences, part colors, local part geometry, and scene-level part statistics. The split covers five question types: part existence, part counting, part color, part spatial relation, and cross-object questions. Together, these question types evaluate whether a model can recognize object parts, reason about missing parts, count repeated components, describe local appearance, infer intra-object spatial relations, and aggregate part information across the scene.

D Additional Qualitative Comparisons

We provide additional qualitative examples to complement the quantitative results in the main paper. The examples are grouped into two figures, covering visual question answering and referring segmentation across multiple evaluation datasets.

Figure 5 shows additional question answering examples from ScanQA and ScenePart-QA datasets. Each example includes the input scene, the ground-truth answer, the answer predicted by PAR3D, and the answer predicted by 3D-LLaVA. These cases illustrate that PAR3D can answer questions that require understanding object attributes, object parts, spatial relationships, and scene-level context. Compared to 3D-LLaVA, PAR3D provides more accurate responses on questions requiring fine-grained 3D scene understanding.

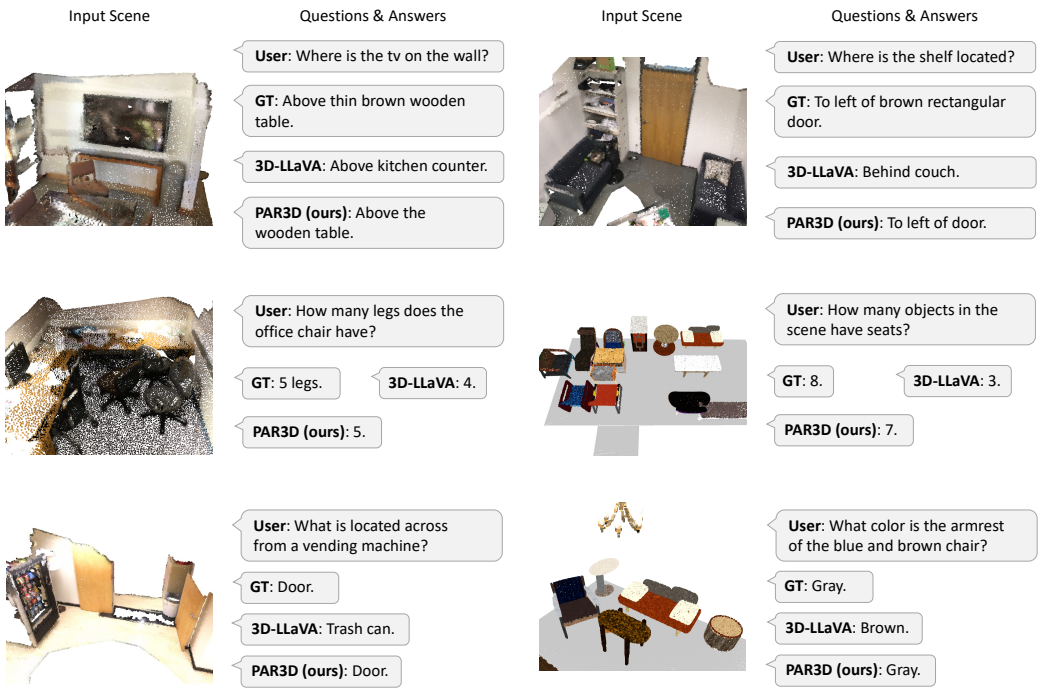

Figure 5: **Additional Qualitative Comparisons on Visual Question Answering.** Each example includes the input scene, the ground-truth answer, the prediction of 3D-LLaVA, and the prediction of PAR3D. PAR3D produces more accurate answers across representative examples from multiple datasets.

Figure 6 shows additional referring segmentation examples from ScanRefer, Multi3DRefer, and ScenePart-Seg datasets. Each example includes the input scene, the ground-truth mask, the mask predicted by PAR3D, and the mask predicted by 3D-LLaVA. The results show that PAR3D achieves more accurate segmentation of language-specified targets in 3D scenes across multiple granularities and benchmark settings.

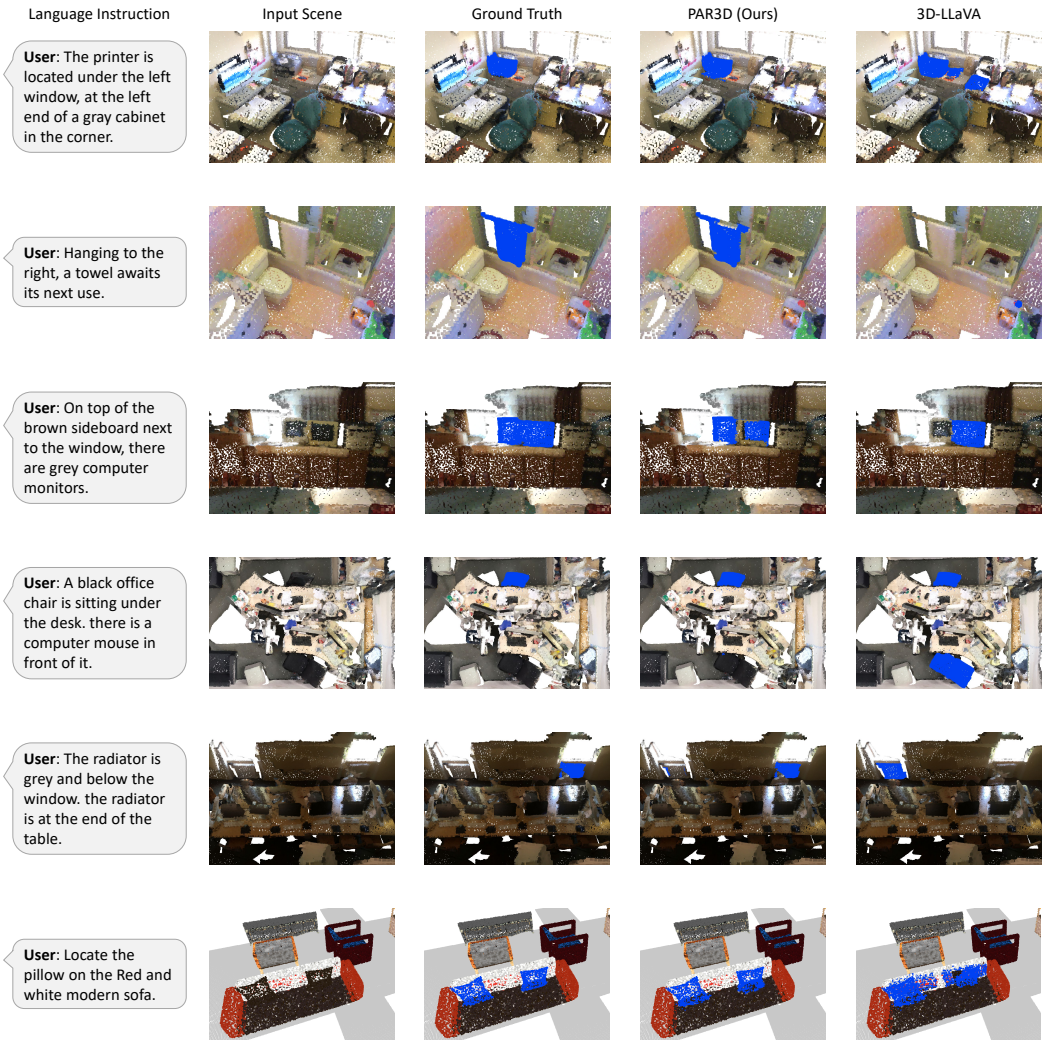

Figure 6: **Additional Qualitative Comparisons on Referring Segmentation.** Each example includes the input scene, the ground-truth mask, the prediction of PAR3D, and the prediction of 3D-LLaVA. The target mask is highlighted in blue, regardless of whether the target corresponds to an object or a part. PAR3D achieves more accurate segmentation across representative examples from multiple datasets.