PAR3D: A UNIFIED 3D-MLLM WITH PART-AWARE REPRESENTATION FOR SCENE UNDERSTANDING

Anonymous authors
Paper under double-blind review

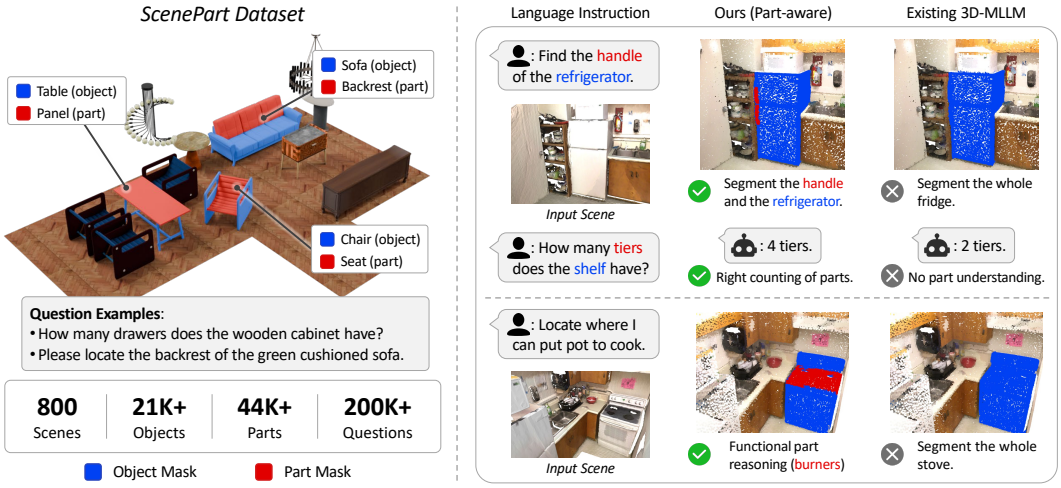

Figure 1: We propose PAR3D, a unified 3D-MLLM with part-aware representation, together with ScenePart dataset. *Left:* ScenePart provides fine-grained object-part annotations and language instructions for 3D scenes. *Right:* PAR3D enables part-aware understanding across QA, segmentation, and reasoning, 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, while also achieving strong performance across object-level vision-language tasks.

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

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with large language models, enabling a unified interface for 3D scene perception and spatial reasoning (Ma et al., 2024; Zha et al., 2025; Huang et al., 2024b;a). 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. 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 (Huang et al., 2024a; Zemskova & Yudin, 2025), language alignment (Hong et al., 2023; Fu et al., 2024), and task supervision (Deng et al., 2025; He et al., 2024). 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 (Dai et al., 2017; Wald et al., 2019; Baruch et al., 2021) 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 (Huang et al., 2024b; Deng et al., 2025). 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 (Deng et al., 2025; Huang et al., 2025b).

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, and generalizes from synthetic part supervision to real-world ScanNet scenes.

2 Related Works

2.1 3D Vision-Language Models for 3D Scenes

3D vision-language research connects natural language with 3D scenes through tasks such as grounding, captioning, and question answering. Early benchmarks establish these task formula-

tions, including ScanRefer and ReferIt3D for 3D grounding (Chen et al., 2020; Achlioptas et al., 2020), Scan2Cap for dense captioning (Chen et al., 2021), and ScanQA and SQA3D for question answering (Azuma et al., 2022; Ma et al., 2023). Subsequent models improve 3D-language alignment through vision-language pretraining, cross-modal interaction, and joint modeling of related tasks (Zhu et al., 2023; Chen et al., 2023a;b; Wu et al., 2024).

Recent 3D-MLLMs further integrate LLMs with complete 3D scenes for instruction following, open-ended reasoning, and unified vision-language understanding (Hong et al., 2023; Huang et al., 2024a; Chen et al., 2024a; Huang et al., 2025a). They explore diverse scene representations, including object-centric tokens, direct point-cloud features, and multi-view or position-aware visual tokens (Huang et al., 2024a; Deng et al., 2025; Zhu et al., 2025; Zheng et al., 2025), and increasingly support grounding, captioning, question answering, and reasoning within a unified framework (Zhu et al., 2024; Huang et al., 2025b). Despite these advances, language understanding in complete 3D scenes remains largely object-centric, with grounding and reasoning primarily targeting whole objects. In contrast, our work explicitly models the object–part hierarchy in complete 3D scenes, extending unified 3D-MLLMs toward part-aware reasoning and dense grounding.

2.2 3D Part Perception

3D part perception has evolved from supervised part segmentation toward more generalizable and task-aware understanding of object components. Early studies rely on part-annotated datasets such as ShapeNetPart (Chang et al., 2015; Yi et al., 2016) and PartNet (Mo et al., 2019), with point-cloud networks (Qi et al., 2017; Li et al., 2018; Wang et al., 2019; Zhao et al., 2021) predicting predefined semantic part labels. Recent works extend part perception beyond fixed taxonomies through foundation-model transfer and large-scale supervision (Zhou et al., 2024; Yang et al., 2024; Umam et al., 2024; Ma et al., 2025b;a), enabling open-world, text-queryable, or promptable part segmentation. Beyond mask prediction, PartField (Liu et al., 2025) and PA3FF (Chen et al., 2026) learn part-aware 3D representations for decomposition, correspondence, and functional manipulation, while Part-X-MLLM (Wang et al., 2026) further supports language-conditioned part understanding of individual 3D objects. These methods predominantly study parts within isolated 3D objects or shapes. In contrast, we study object parts in complete 3D scenes, explicitly model their relationships with host objects and scene context, and incorporate the object–part hierarchy into a unified 3D-MLLM for fine-grained language reasoning and point-level 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 (Deng et al., 2025), 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:

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

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:

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

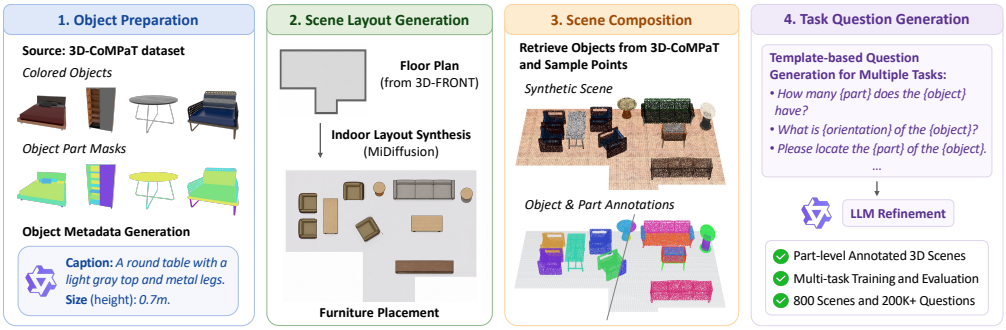

Figure 2: **ScenePart Data Construction Pipeline.** ScenePart composes part-annotated 3D objects into synthesized indoor layouts, producing object- and part-level mask annotations in 3D scenes and multi-task language instructions for training and evaluating part-aware 3D-MLLMs.

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

To enable referring segmentation, 3D-LLaVA uses a special token [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 from existing 3D vision-language data: indoor-scene datasets provide object-level annotations (Dai et al., 2017; Baruch et al., 2021) and language tasks (Azuma et al., 2022; Chen et al., 2020), whereas part-annotated datasets mainly focus on isolated objects (Mo et al., 2019; Li et al., 2022). 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 a scene graph recording 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 QA and referring segmentation instructions. For evaluation, we further construct two test splits: **ScenePart-QA** evaluates part-aware question answering with diverse question types, while **ScenePart-Seg** evaluates referring segmentation across objects and parts at different granularities.

We construct ScenePart in four steps, as illustrated in Figure 2. First, we preprocess part-annotated 3D assets from 3D-CoMPaT (Li et al., 2022; Slim et al., 2025) by filtering object models and normalizing their sizes. We use Qwen3-VL-8B (Bai et al., 2025) to estimate object scales and generate object descriptions that are later used for language annotation. Second, we generate indoor layouts using MiDiffusion (Hu et al., 2026) on floor plans from 3D-FRONT (Fu et al., 2021a), 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 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

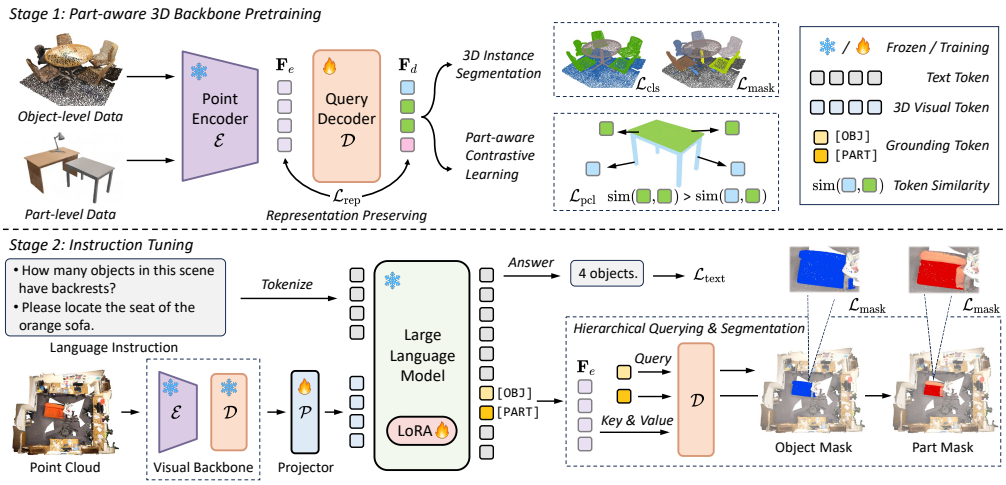

Figure 3: **Overall Framework of PAR3D.** PAR3D is trained with a two-stage scheme. Stage 1 adapts the 3D visual backbone with object- and part-level supervision through instance segmentation, part-aware contrastive learning, and representation-preserving regularization. Stage 2 performs instruction 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].

representation, we instantiate the 3D encoder $\mathcal{E}$ with a pretrained Point Transformer model (Wu et al., 2025; Zhang et al., 2026). 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 comprises a frozen pretrained encoder $\mathcal{E}$ and a trainable query decoder $\mathcal{D}$. The decoder features $\mathbf{F}_d$ are used to construct visual tokens. However, task adaptation can bias $\mathbf{F}_d$ toward segmentation-specific representations and weaken the general semantics inherited from the encoder. We therefore employ two regularization objectives, part-aware contrastive learning and representation-preserving self-distillation, to enhance part-level structure while preserving encoder semantics.

Part-aware contrastive learning. To encode fine-grained part structure, we encourage features within the same part to be compact while separating those from different parts. Following supervised contrastive learning (Khosla et al., 2020), we use ScenePart part masks to construct superpoint-level positive and negative sets for $\mathbf{F}_d$. For an anchor $\mathbf{f}_i^d$, $\mathcal{P}(i)$ contains other features in the same part, whereas $\mathcal{N}(i)$ contains features from different parts. The resulting part-aware contrastive loss follows an InfoNCE objective (Oord et al., 2018):

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

where τ is the temperature and $\text{sim}(\cdot,\cdot)$ 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 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:

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

where M is the number of superpoints and $\text{sg}(\cdot)$ 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 [SEG] to represent referred targets of all granularities (Deng et al., 2025; Huang et al., 2025b). 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 [OBJ] to represent the referred object. For part-level grounding, it generates [OBJ] followed by [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 [OBJ] and [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, the model is supervised to generate [OBJ] and predict the corresponding object mask for object-level expressions. For part-level expressions, it is supervised to generate both [OBJ] and [PART] and predict the host-object and target-part masks. Joint token and mask supervision preserves object–part structure within a unified grounding interface.

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.

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:

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

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

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

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 (Chen et al., 2020), Nr3D (Achlioptas et al., 2020), Multi3DRefer (Zhang et al., 2023), ScanQA (Azuma et al., 2022), SQA3D (Ma et al., 2023), and

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 (Azuma et al., 2022)	-	-	64.9	10.1	13.1	33.3	46.6	-	-	-	-	-
3D-VisTA (Zhu et al., 2023)	-	-	69.6	10.4	13.9	**45.7**	48.5	-	61.6	34.1	26.8	55.0
Scan2Cap (Chen et al., 2021)	-	-	-	-	-	-	41.0	-	39.1	23.3	22.0	44.8
UniT3D (Chen et al., 2023a)	-	-	-	-	-	-	-	-	46.7	27.2	21.9	46.0
Vote2Cap-DETR (Chen et al., 2023b)	-	-	-	-	-	-	-	-	61.8	34.5	26.2	54.4
M3DRef-CLIP (Zhang et al., 2023)	35.7	32.6	-	-	-	-	-	-	-	-	-	-
3D-STMN (Wu et al., 2024)	39.5	-	-	-	-	-	-	-	-	-	-	-
Finetuned 3D-MLLMs:												
3D-LLM (Hong et al., 2023)	-	-	69.4	12.0	14.5	35.7	-	-	-	-	-	-
Scene-LLM (Fu et al., 2024)	-	-	80.0	12.0	16.8	40.0	54.2	-	-	-	-	-
LL3DA (Chen et al., 2024a)	-	-	76.8	13.5	15.9	37.3	-	-	65.2	36.8	26.0	55.1
SegPoint (He et al., 2024)	41.7	36.1	-	-	-	-	-	-	-	-	-	-
Generalist 3D-MLLMs:												
LEO (Huang et al., 2024b)	-	-	101.4	13.2	20.0	49.2	50.0	52.4	72.4	**38.2**	27.9	**58.1**
Scene-LLM (Fu et al., 2024)	-	-	80.0	11.7	15.8	35.9	53.6	-	-	-	-	-
Reason3D (Huang et al., 2025b)	42.0	-	73.5	12.1	15.1	37.4	-	-	-	-	-	-
Chat-Scene (Huang et al., 2024a)	-	-	87.7	14.3	18.0	41.6	54.6	**57.5**	77.2	36.4	**28.0**	**58.1**
Grounded 3D-LLM (Chen et al., 2024b)	-	-	72.7	13.4	-	-	-	-	70.6	35.5	-	-
3DGraphLLM (Zemskova & Yudin, 2025)	-	-	88.8	15.9	-	-	**55.9**	-	81.0	36.5	-	-
3D-LLaVA (Deng et al., 2025)	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

Scan2Cap (Chen et al., 2021), covering object-level referring segmentation, question answering, and captioning. ScenePart further introduces part-aware QA and referring segmentation instructions.

All instruction-tuning tasks use the standard autoregressive loss $\mathcal{L}_{\text{text}}$ to supervise textual responses, including the generation of $[\text{OBJ}]$ for object-level targets and both $[\text{OBJ}]$ and $[\text{PART}]$ for part-level targets. Grounding tasks additionally use $\mathcal{L}_{\text{mask}}$ to supervise the corresponding object and part masks as described in Section 3.4. 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 Datasets and Metrics

Datasets. We evaluate on the ScenePart test splits, a real-world ScanNet part test set, and standard 3D vision-language benchmarks. ScenePart-Seg and ScenePart-QA, introduced in Section 3.2, evaluate multi-granularity referring segmentation and part-aware question answering, respectively. The real-world test set contains 114 referring cases with manually annotated target masks from 52 ScanNet validation scenes unused during training. For object-level evaluation, we use ScanRefer (Chen et al., 2020) and Multi3DRefer (Zhang et al., 2023) for referring segmentation, ScanQA (Azuma et al., 2022) and SQA3D (Ma et al., 2023) for question answering, and Scan2Cap (Chen et al., 2021) for dense captioning.

Metrics. For referring segmentation, we report mean Intersection-over-Union (mIoU) between predicted and ground-truth masks. On ScenePart-Seg and the real-world test set, we additionally report Acc@0.5, the percentage of predictions with an IoU above 50%. For language-generation tasks, we report CIDEr (C), BLEU-4 (B-4), METEOR (M), and ROUGE-L (R-L), while SQA3D uses exact-match accuracy (EM) and refined exact-match accuracy (EM-R). Higher values indicate better performance; detailed metric definitions are provided in Appendix E.

4.2 Comparison with State-of-the-Art Methods

We compare PAR3D with existing approaches across multiple 3D vision-language tasks, covering both object-level and part-level evaluation. Our comparison focuses on 3D-MLLMs that directly take 3D point clouds as visual input and support one or more tasks, including 3D segmentation, localization, and language generation. 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

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 (Deng et al., 2025)	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 (Huang et al., 2025b)	25.5	13.8	7.4	2.3	3.9	0.8	9.6	4.0	-	-	-	-
Chat-Scene (Huang et al., 2024a)	37.1	38.0	11.3	6.5	5.3	1.8	14.1	10.9	35.1	0.0	7.9	27.9
3DGraphLLM (Zemskova & Yudin, 2025)	10.3	10.5	4.1	1.5	1.6	0.5	4.3	2.9	1.7	0.0	3.0	5.6
Grounded 3D-LLM (Chen et al., 2024b)	31.9	32.8	11.0	7.1	4.9	2.1	12.7	10.2	36.9	0.0	8.2	30.9
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**

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-aware understanding and grounding.

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. The ScanQA results of LEO (Huang et al., 2024b) 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, including 3D-LLaVA, Reason3D, Chat-Scene, 3DGraphLLM, and Grounded 3D-LLM. Further discussion of the evaluated 3D-MLLM baselines in the context of part-level scene understanding is provided in Appendix F. For ScenePart-Seg, we report mIoU and Acc@0.5 for object targets, coarse parts (*e.g.*, door of a refrigerator), and fine-grained parts (*e.g.*, handle of a refrigerator), together with overall results across all targets. For ScenePart-QA, we report standard language-generation metrics. PAR3D consistently outperforms all compared 3D-MLLMs across segmentation granularities and question-answering metrics.

To further evaluate generalization to real-world scans, Table 3 reports results on our manually annotated part-level referring segmentation test set constructed from ScanNet scenes. PAR3D achieves 39.4% mIoU and 32.5% Acc@0.5, surpassing the strongest baseline for each metric by absolute gains of 6.2% and 7.1%, respectively. Together, the results on synthetic ScenePart scenes and real-world ScanNet scans demonstrate that PAR3D learns effective part-aware representations that generalize across domains, although its part-level supervision is provided only by synthetic data during training.

Table 3: **Quantitative Comparison on Real-World Part Referring Segmentation.**

Methods	mIoU↑	Acc@0.5↑
3D-LLaVA (Deng et al., 2025)	31.2	25.4
Reason3D (Huang et al., 2025b)	7.1	2.6
Chat-Scene (Huang et al., 2024a)	33.2	22.8
3DGraphLLM (Zemskova & Yudin, 2025)	4.0	2.6
Grounded 3D-LLM (Chen et al., 2024b)	29.6	21.9
PAR3D (ours)	**39.4**	**32.5**

Qualitative results. Figure 4 presents representative examples on real ScanNet scenes with part-aware instructions for referring segmentation and QA. In segmentation, blue and red masks indicate object- and part-level predictions, respectively. PAR3D segments parts such as the pillow and bucket handle together with their host objects, while 3D-LLaVA often produces coarser object masks or misses the referred part. In part-aware question answering, PAR3D correctly counts chairs with

armrests. These examples demonstrate PAR3D's ability to handle fine-grained object-part understanding across referring segmentation and question answering. Additional qualitative comparisons are provided in Appendix H.

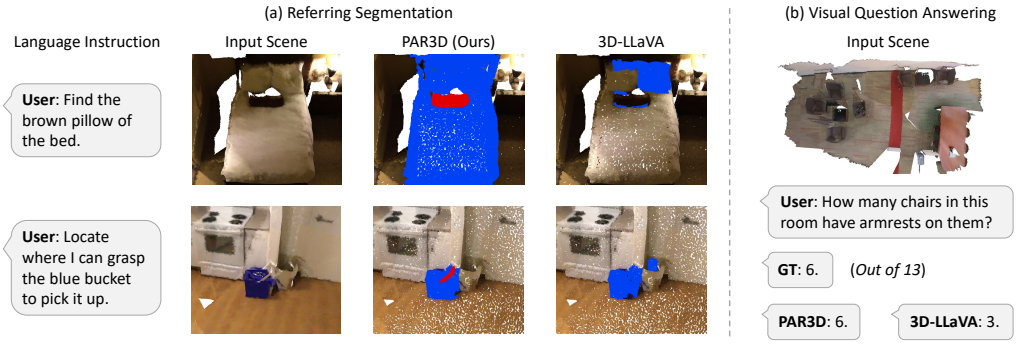

Figure 4: **Qualitative Examples of PAR3D.** We present examples on (a) referring segmentation and (b) visual question answering for part-aware 3D scene understanding.

4.3 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 4, 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 4: **Ablation Studies on Referring Segmentation.**

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. Evaluation on the manually annotated ScanNet part test set further provides evidence that PAR3D generalizes from synthetic part supervision to real-world scans. We hope this work encourages future unified 3D-MLLMs toward fine-grained perception, reasoning, and grounding over objects and their parts.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to assist in drafting selected appendix passages, and to support the construction of the synthetic ScenePart dataset. All AI-assisted text was reviewed and revised by the authors, who verified the associated factual claims, citations, and reported results.

For ScenePart construction, Qwen3-VL-8B (Bai et al., 2025) was used to estimate object scales, generate object descriptions, and refine template-generated language annotations, while MiDiffusion (Hu et al., 2026) was used to generate indoor scene layouts. These components are described in Section 3.2 and Appendix B. We reviewed all AI-assisted work and take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 describes the model architecture, part-aware representation learning objectives, hierarchical segmentation query generation, and training strategy. Detailed implementation settings, training hyperparameters, compute resources, and training time are provided in Appendix A. The construction of ScenePart and its evaluation splits is documented in Section 3.2 and Appendices B and C, while the construction of the real-world ScanNet part test set is described in Appendix D. The datasets, evaluation metrics, and baseline evaluation protocols used in the experiments are specified in Section 4, Appendix E, and Appendix F, respectively.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pp. 422–440. Springer, 2020.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 19107–19117. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01854. URL https://doi.org/10.1109/CVPR52688.2022.01854.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *ArXiv preprint*, abs/2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *ArXiv preprint*, abs/2111.08897, 2021. URL https://arxiv.org/abs/2111.08897.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *ArXiv preprint*, abs/1512.03012, 2015. URL https://arxiv.org/abs/1512.03012.

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

Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in RGB-D scans. In *IEEE Conference on Computer Vision and Pattern*

Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 3193–3203. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00321. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Scan2Cap_Context-Aware_Dense_Captioning_in_RGB-D_Scans_CVPR_2021_paper.html.

Dave Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X. Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 18063–18073. IEEE, 2023a. doi: 10.1109/ICCV51070.2023.01660. URL https://doi.org/10.1109/ICCV51070.2023.01660.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 11124–11133. IEEE, 2023b. doi: 10.1109/CVPR52729.2023.01070. URL https://doi.org/10.1109/CVPR52729.2023.01070.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26418–26428. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02496. URL https://doi.org/10.1109/CVPR52733.2024.02496.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *ArXiv preprint*, abs/2405.10370, 2024b. URL https://arxiv.org/abs/2405.10370.

Yue Chen, Muqing Jiang, Kaifeng Zheng, Jiaqi Liang, Chenrui Tie, Haoran Lu, Ruihai Wu, and Hao Dong. Pa3ff: Learning part-aware dense 3d feature field for generalizable articulated object manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 50604–50626, 2026.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2432–2443. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.261. URL https://doi.org/10.1109/CVPR.2017.261.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3772–3782, 2025.

Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 10913–10922. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.01075. URL https://doi.org/10.1109/ICCV48922.2021.01075.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129 (12):3313–3337, 2021b.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *ArXiv preprint*, abs/2403.11401, 2024. URL https://arxiv.org/abs/2403.11401.

Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pp. 349–367. Springer, 2024.

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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/413885e70482b95dcbeeddc1daf39177-Abstract-Conference.html.

Siyi Hu, Diego Martin Arroyo, Stephanie Debats, Fabian Manhardt, Luca Carlone, and Federico Tombari. Mixed diffusion for 3d indoor scene synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1262–1272, 2026.

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. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024a.

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. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=V4qV08Vk6S.

Jiangyong Huang, Xiaojian Ma, Xiongkun Linghu, Junchao He, Qing Li, Song-Chun Zhu, Yixin Chen, Baoxiong Jia, and Siyuan Huang. Leo-vl: Efficient scene representation for scalable 3d vision-language learning. *arXiv preprint arXiv:2506.09935*, 2025a.

Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In *2025 International Conference on 3D Vision (3DV)*, pp. 1177–1186. IEEE, 2025b.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 828–838, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html.

Yuchen Li, Ujjwal Upadhyay, Habib Slim, Ahmed Abdelreheem, Arpit Prajapati, Suhail Pothigara, Peter Wonka, and Mohamed Elhoseiny. 3d compat: Composition of materials on parts of 3d things. In *European conference on computer vision*, pp. 110–127. Springer, 2022.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9704–9715. IEEE, 2025.

Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-sam: Native 3d part segmentation. *ArXiv preprint*, abs/2509.06784, 2025a. URL https://arxiv.org/abs/2509.06784.

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

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=IDJx97BC38.

Ziqi Ma, Yisong Yue, and Georgia Gkioxari. Find any part in 3d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7818–7827, 2025b.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 909–918. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00100. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Mo_PartNet_A_Large-Scale_Benchmark_for_Fine-Grained_and_Hierarchical_Part-Level_3D_CVPR_2019_paper.html.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv preprint*, abs/1807.03748, 2018. URL https://arxiv.org/abs/1807.03748.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5099–5108, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html.

Habib Slim, Xiang Li, Yuchen Li, Mahmoud Ahmed, Mohamed Ayman, Ujjwal Upadhyay, Ahmed Abdelreheem, Arpit Prajapati, Suhail Pothigara, Peter Wonka, et al. 3dcompat++: An improved large-scale 3d vision dataset for compositional recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. Partdistill: 3d shape part segmentation by vision-language model distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 3470–3479. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00333. URL https://doi.org/10.1109/CVPR52733.2024.00333.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3d object instance re-localization in changing indoor environments. In *2019 IEEE/CVF International*

Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 7657–7666. IEEE, 2019. doi: 10.1109/ICCV.2019.00775. URL https://doi.org/10.1109/ICCV.2019.00775.

Chunshi Wang, Junliang Ye, Yunhan Yang, Yang Li, Zizhuo Lin, Jun Zhu, Zhuo Chen, Yawei Luo, and Chunchao Guo. Part-x-mllm: Part-aware 3d multimodal large language model. In *International Conference on Learning Representations*, volume 2026, pp. 150386–150417, 2026.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12, 2019.

Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 5940–5948. AAAI Press, 2024. doi: 10.1609/AAAI.V38I6.28408. URL https://doi.org/10.1609/aaai.v38i6.28408.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22193–22204, 2025.

Yunhan Yang, Yukun Huang, Yuan-Chen Guo, Liangjun Lu, Xiaoyang Wu, Edmund Y Lam, Yan-Pei Cao, and Xihui Liu. Sampart3d: Segment any part in 3d objects. *ArXiv preprint*, abs/2411.07184, 2024. URL https://arxiv.org/abs/2411.07184.

Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

Tatiana Zemskova and Dmitry Yudin. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8885–8895, 2025.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *ArXiv preprint*, abs/2504.05786, 2025. URL https://arxiv.org/abs/2504.05786.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15179. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01397. URL https://doi.org/10.1109/ICCV51070.2023.01397.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds. *ArXiv preprint*, abs/2603.03283, 2026. URL https://arxiv.org/abs/2603.03283.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 16239–16248. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01595. URL https://doi.org/10.1109/ICCV48922.2021.01595.

Duo Zheng, Shijia Huang, and Liwei 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*, pp. 8995–9006, 2025.

Yuchen Zhou, Jiayuan Gu, Tung Yen Chiang, Fanbo Xiang, and Hao Su. Point-sam: Promptable 3d segmentation model for point clouds. *ArXiv preprint*, abs/2406.17741, 2024. URL https://arxiv.org/abs/2406.17741.

Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision*, pp. 151–168. Springer, 2024.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui 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*, pp. 4295–4305, 2025.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2899–2909. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00272. URL https://doi.org/10.1109/ICCV51070.2023.00272.

A Implementation Details

The LLM backbone of our work is LLaVA-1.5-7B (Liu et al., 2024), following the setting of 3D-LLaVA (Deng et al., 2025). Specifically, the original 3D encoder is replaced with the pretrained Utonia encoder (Zhang et al., 2026). Unless otherwise specified, we use the same preprocessing protocol as 3D-LLaVA.

The training proceeds in two stages. In Stage 1, we train the query decoder on ScanNet200 (Dai et al., 2017) and ScenePart while keeping the pretrained 3D encoder frozen. This stage is trained for 256 epochs using AdamW with an initial learning rate of 3×10^{-4} and polynomial learning-rate decay. The batch size is set to 64, and the loss weights are set to $\lambda_{\text{pcl}} = 0.1$ and $\lambda_{\text{rep}} = 0.6$ for the part-aware contrastive loss and representation-preserving loss, respectively.

In Stage 2, we perform LoRA-based instruction tuning for 2 epochs on the instruction datasets described in Sec. 3.5. We keep the main LLM parameters frozen and optimize the projector MLP $\mathcal{P}$, the segmentation MLP ϕ, and LoRA parameters with AdamW, using an initial learning rate of 2×10^{-4} and a cosine annealing schedule. The batch size is set to 8. We use LoRA with rank 32 and scaling factor 64. For grounding tasks, the mask loss weight is set to $\lambda_{\text{mask}} = 0.1$.

All experiments are conducted on 4 NVIDIA A100 GPUs with 40GB memory. Stage 1 training takes approximately 10 hours, and Stage 2 instruction tuning takes approximately 24 hours.

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 (Li et al., 2022; Slim et al., 2025). 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 (Bai et al., 2025) 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 (Hu et al., 2026), a diffusion-based scene synthesis model trained on 3D-FRONT (Fu et al., 2021a). 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 (Fu et al., 2021b) 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 5: **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 6: **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 Real-World Part Annotation on ScanNet

To evaluate part-level grounding in real-world scans, we construct a manually annotated test set from the ScanNet validation split; none of the selected scenes are used during training. We first randomly sample 60 scenes and then screen them for annotation suitability. Eight scenes are excluded because severe reconstruction artifacts make their part boundaries difficult to identify reliably or because they contain too few objects with clearly identifiable parts. From the remaining 52 scenes, we construct 114 part-level referring-segmentation samples. Each sample consists of a referring expression and a corresponding point-level target-part mask. We annotate the target masks directly on the ScanNet point clouds using a customized version of SegGroup.annotator[1]. Unlike ScenePart-Seg, which distinguishes coarse- and fine-grained parts according to the source part hierarchy, this real-world test set pools all part targets into a single evaluation split without imposing a coarse/fine distinction.

E 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 on ScenePart-Seg, the ScanNet part test set, ScanRefer (Chen et al., 2020), and Multi3DRefer (Zhang et al., 2023), we report mean intersection-over-union (mIoU) between the predicted and ground-truth masks. For ScenePart-Seg and the ScanNet part test set, 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 (Azuma et al., 2022), we report CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and BLEU-4 (Papineni et al., 2002). 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 (Ma et al., 2023), we report EM and EM-R (Huang et al., 2024b). 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.

[1] https://github.com/AnTao97/SegGroup.annotator

Dense captioning metrics. For dense captioning on Scan2Cap (Chen et al., 2021), 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.

F Part-Level Evaluation of 3D-MLLM Baselines

Comparison scope. Our part-level evaluation follows the native 3D setting considered in this work: a complete scene is represented as a point cloud, a language instruction specifies the target or question, and grounding is evaluated as a point-level mask in the same 3D scene. Accordingly, we compare with 3D-MLLMs whose released inference pipelines directly take 3D point clouds as visual input and provide localized predictions that can be evaluated under this protocol.

Multi-view based 3D-MLLMs, such as LLaVA-3D (Zhu et al., 2025) and Video-3D LLM (Zheng et al., 2025), adopt a different observation and grounding interface and do not directly produce point-level masks under the same evaluation protocol. We therefore discuss these methods in the related work rather than include them in the quantitative part-grounding comparison.

Part-level behavior of existing 3D-MLLMs. The evaluated baselines cover different 3D grounding interfaces. Chat-Scene (Huang et al., 2024a), 3DGraphLLM (Zemskova & Yudin, 2025), and Grounded 3D-LLM (Chen et al., 2024b) ground language through object-centric scene representations and produce localized object predictions. Their grounding granularity is therefore determined by the upstream object segmentation, which does not explicitly expose internal part regions to the language model. In contrast, 3D-LLaVA (Deng et al., 2025) and Reason3D (Huang et al., 2025b) directly decode language-conditioned point-level masks. Although their mask decoders are not restricted to predefined object proposals, these models lack part-level supervision and explicit object–part modeling.

As shown in Table 2, part-level grounding remains challenging across different grounding interfaces. 3D-LLaVA decreases from 29.0% mIoU on object targets to 9.0% and 4.2% mIoU on coarse- and fine-part targets, respectively. Similar trends are observed for Chat-Scene (37.1%, 11.3%, and 5.3%) and Grounded 3D-LLM (31.9%, 11.0%, and 4.9%) on object, coarse-part, and fine-part targets, respectively. The degradation becomes particularly pronounced for fine-grained parts, indicating that object-level grounding capability does not directly translate into localization of internal object components.

The difficulty also persists on the manually annotated ScanNet part-level test set in Table 3. On real scans, the stronger existing baselines obtain 31.2% mIoU for 3D-LLaVA, 33.2% for Chat-Scene, and 29.6% for Grounded 3D-LLM, compared with 39.4% for PAR3D. Although the performance gap is smaller than on ScenePart-Seg, the real-scene results show that fine-grained part grounding remains challenging beyond the synthetic ScenePart setting. These results highlight the value of part-level supervision and explicit object–part modeling when extending 3D-MLLMs from object-level to part-level scene understanding.

Evaluation protocol and adaptations. All baselines are evaluated using their released checkpoints and original inference code. We only make the input/output adaptations required to express the ScenePart tasks through their released interfaces. All segmentation predictions are ultimately represented as point masks and evaluated against the same ground-truth target masks using the metrics described in Appendix E. For ScenePart-QA, the generated textual responses are evaluated against the same reference answers using the language-generation metrics described in Appendix E.

For Chat-Scene, 3DGraphLLM, and Grounded 3D-LLM, we use the object mask associated with the model's grounded output as its predicted mask and directly compare it with the target point mask. This preserves the native grounding behavior of each model. For Chat-Scene and 3DGraphLLM, we additionally format the ScenePart referring instructions according to the referential input format expected by their released inference interfaces, while preserving the target semantics. For 3D-LLaVA and Reason3D, we directly use the point-level masks produced by their native mask decoders. Rea-

son3D is evaluated only on ScenePart-Seg because its released inference pipeline does not produce valid open-ended textual answers under the ScenePart-QA protocol.

G LIMITATIONS

Although PAR3D improves part-aware 3D scene understanding, several limitations remain. Part-level supervision during training is derived entirely from synthetic scenes, and the manually annotated real-world test set remains limited in scale. ScenePart is also constrained by the object and part categories available in its source assets. Future work may incorporate more diverse real-world part supervision, extend to open-vocabulary part categories, and study embodied scenarios requiring more complex object–part reasoning and interaction.

H 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.

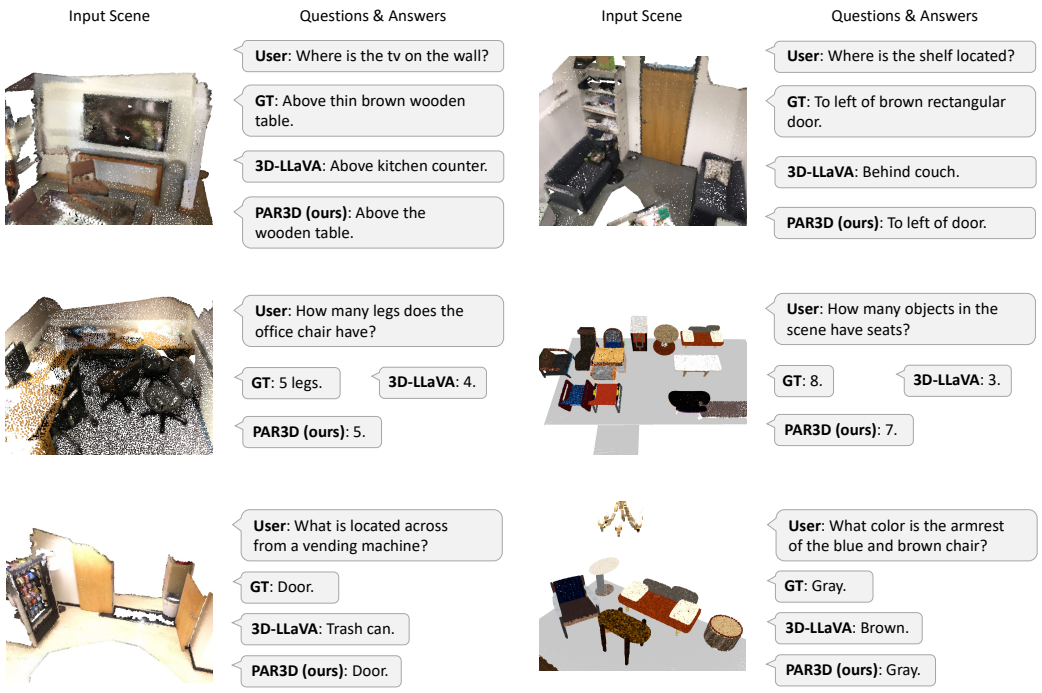

Figure 5: **Additional Qualitative Comparisons on Visual Question Answering.** Each example includes the input scene, the ground-truth answer, the prediction of 3D-LLaVA, and the prediction of PAR3D. PAR3D produces more accurate answers across representative examples from multiple datasets.

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.

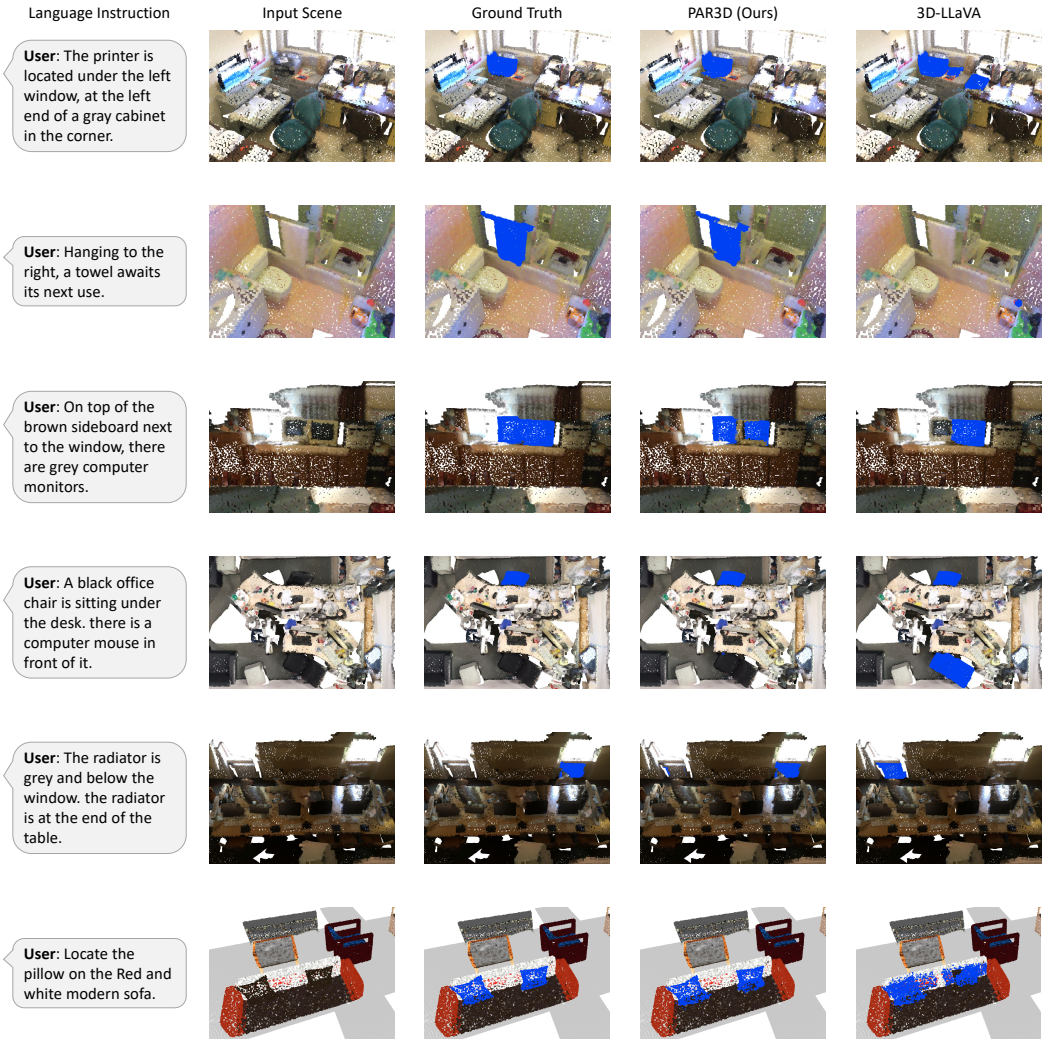

Figure 6: **Additional Qualitative Comparisons on Referring Segmentation.** Each example includes the input scene, the ground-truth mask, the prediction of PAR3D, and the prediction of 3D-LLaVA. The target mask is highlighted in blue, regardless of whether the target corresponds to an object or a part.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