Ground3D-LMM: Fine-Grained 3D Point Grounding and Spatial Reasoning with LMM

Amol Harsh[1], Zongyan Han[1]✉, Jean Lahoud[1], Ye Liu[2], Rao Muhammad Anwer[1], Hisham Cholakkal[1], Salman Khan[1], and Fahad Shahbaz Khan[1,3]

[1] Mohamed bin Zayed University of Artificial Intelligence
{amol.harsh, zongyan.han, jean.lahoud, rao.anwer, hisham.cholakkal, salman.khan, fahad.khan}@mbzuai.ac.ae
[2] The Hong Kong Polytechnic University
coco.ye.liu@connect.polyu.hk
[3] Linköping University

https://github.com/AmolHarsh/ground3d-lmm

Abstract. Natural-language queries about 3D environments become actionable when responses are verifiable and metric. Verifiability requires explicit grounding to the referred 3D region, while metric answers report physical measurements in real-world units (e.g., size, thickness, clearance, and distance). Existing 3D large multimodal models (LMMs) approaches remain limited: conversational systems typically respond without explicit 3D grounding, while 3D grounding models are not designed for interactive, metric-aware dialogue. In this paper, we present **Ground3D-LMM**, a unified model that takes a point cloud and an optional RGB image as input and supports 3D spatial conversation with *(i)* point-grounded responses and *(ii)* metric numeric outputs at both object and part granularity, including multi-object queries. To evaluate this intersection of grounding and measurement, we define the *3D Grounded Measurement* task, which requires predicting the referred 3D region and the corresponding metric quantities in real-world units. We introduce a large-scale dataset built on ScanNet and ScanNet++ datasets with dense object and part annotations and roughly 2.5M question-answer pairs spanning eight tasks, along with a manually verified test set. Extensive experiments on multiple datasets and tasks show that our proposed Ground3D-LMM model provides a strong baseline for grounded, metric-aware 3D conversational understanding. Our dataset and model are publicly available.

1 Introduction

Natural interaction with a 3D environment often requires two capabilities. First, the system must verify what it is talking about by grounding language to the correct region in the scene. Second, it must be able to provide metric measurements: answers such as "the chair is larger" are less useful than answers expressed in

✉ Corresponding author.

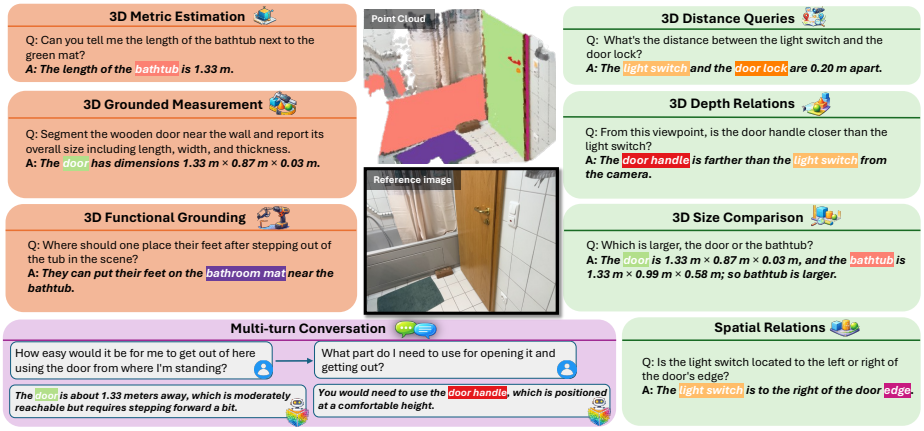


Fig. 1: Ground3D-LMM answers open-vocabulary object/part questions with grounded (3D segmentation) and metric (real-world units) outputs. We show example queries spanning 3D metric estimation, grounded measurement, functional grounding, spatial relations, and multi-object distance/depth/size reasoning. Queries also include multi-turn dialogue with follow-up questions that reuse previously grounded entities; the model outputs a point-level mask when a response contains a `<SEG>` trigger.

meters or centimeters. These requirements appear in various tasks, such as checking clearances in robotics, measuring surfaces in AR/VR, and locating functional parts of objects in assistive agents.

Current 3D vision-language models rarely satisfy both requirements in one system. Some works focus on 3D conversation or question answering, but provide text without explicit point-level grounding. Other works produce 3D masks or boxes from text, but do not support interactive dialogue and are not designed to return metric quantities. Moreover, open-vocabulary part-level reasoning is still underexplored, even though many real requests refer to functional parts (e.g., seat of a chair, handle of a drawer).

In this paper, we propose Ground3D-LMM, a unified model that supports grounded and metric-aware 3D conversation. Given a colored point cloud, a user query, and an optional RGB image, our model predicts a textual response along with a 3D segmentation mask for the referred object or part. Moreover, measurements are returned in metric physical units and support multi-object queries (e.g., comparing two chairs, or measuring the clearance between a desk and a drawer). Figure 1 shows examples of the open-vocabulary object/part queries with grounded and metric outputs. Our key idea is to embed 3D point features as tokens alongside text (and optional image tokens) in an LMM, so the model can reason jointly over language and 3D geometry. When the model generates a special grounding trigger (e.g., `<SEG>`) tied to a phrase, a lightweight segmentation head predicts the corresponding point-level mask.

The key contributions of our work are as follows:

Table 1: Comparison of recent 3D multimodal / conversational models. We compare input modalities and whether a method supports *3D spatial conversation with metric values*, *point-grounded conversational outputs* (text paired with point-level masks), and grounding granularity. "seg" denotes point-level masks and "num" denotes metric numeric outputs in real units.

Year	Venue	Model / Paper	Input			3D Spatial Conv.	Point Grounded Conv. (Text + Seg)	Grounding Granularity		
			RGB	Point Cloud	Scene level			Object (seg / num)	Part (seg / num)	Multi Object
2025	CVPR	Inst3D-LLM [37]	✓	✓	✓	✗	✗	✗	✗	✓
2025	NeurIPS	SD-VLM [8]	✓	✗	✓	✓	✗	✓	✗	✓
2025	NeurIPS	MLLM-For3D [15]	✓	✗	✓	✗	✗	✓	✗	✗
2025	ICCV	Kestrel [2]	✗	✓	✗	✗	✓	✗	✓	✗
2025	CVPR	SeeGround [21]	✗	✓	✓	✗	✗	✗	✗	✗
2024	ICLR	Reason3D [16]	✗	✓	✓	✗	✗	✓	✗	✗
2024	ECCV	SegPoint [13]	✗	✓	✓	✗	✗	✓	✗	✓
2024	ICLR	Open-YOLO 3D [5]	✓	✓	✓	✗	✗	✓	✗	✗
2024	CVPR	Open3DIS [27]	✗	✓	✓	✗	✗	✓	✗	✗
2024	ECCV	PARIS-3D [20]	✗	✓	✗	✗	✓	✗	✓	✗
2023	CVPR	PartSLIP [23]	✗	✓	✗	✗	✗	✗	✓	✗
Ours			✓	✓	✓	✓	✓	✓	✓	✓

– We present **Ground3D-LMM**, a unified point-cloud LMM that produces text and 3D masks in a single conversational interface, and supports metric measurements at object and part granularity.
– We introduce a large-scale dataset with dense 3D object/part annotations and multi-turn conversations designed for grounded and metric reasoning, along with a curated evaluation split with manual verification.
– We formalize *3D Grounded Measurement*, a task that requires predicting a referred 3D region and metric quantities under a single evaluation protocol.
– We conduct extensive experiments on our proposed dataset and relevant subtask benchmarks, demonstrating that Ground3D-LMM consistently achieves superior performance across all tasks.

2 Related Work

We review prior work on 3D language grounding, open-vocabulary 3D segmentation, 3D multimodal LLMs, and metric spatial reasoning. Table 1 summarizes representative recent systems and highlights which components are typically missing for grounded, metric-aware 3D conversation (e.g., point-level grounding, part granularity, and multi-object measurement).

Language grounding in 3D. 3D referring expression comprehension and grounding benchmarks require localizing a target object in a scene, usually with a box or a mask. Benchmarks such as ScanRefer [7] and ReferIt3D [1] established the importance of explicit grounding for verifiable 3D interaction. Recent systems in this direction include Kestrel [2], PARIS-3D [20], SegPoint [13], and SeeGround [21]. Referring segmentation methods such as InstanceRefer [38],

BUTD-DETR [18], EDA [33], and 3D-STMN [32] further improve localization quality, while Multi3DRefer [39] studies multi-object referring expressions. These methods improve localization and grounding quality, but they typically do not support interactive, multi-turn dialogue or metric outputs.

Open-vocabulary 3D segmentation. Recent open-vocabulary methods extend 2D vision-language pre-training to 3D point clouds and enable class-agnostic segmentation. Representative approaches include Open3DIS [27], Open-YOLO 3D [5], PartSLIP [23], as well as UniSeg3D [35]. While these models are strong at producing masks, they are not designed to hold a conversation or provide metric answers with unit consistency, especially at the part level.

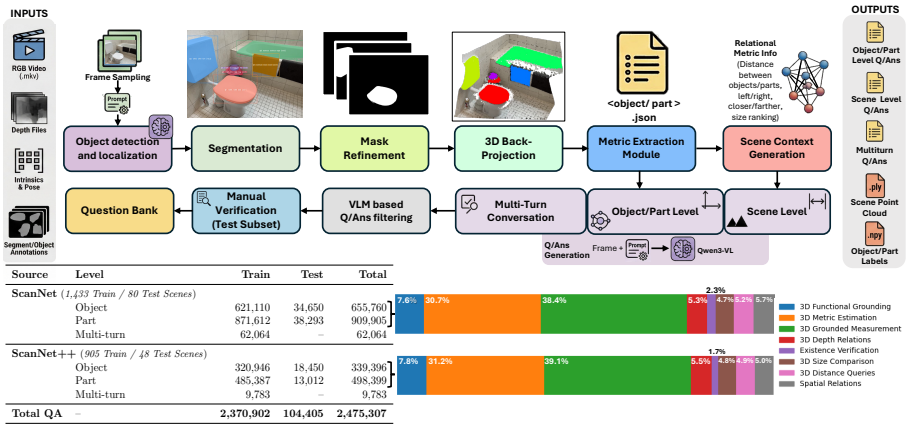
3D question answering and multimodal LLMs. 3D QA datasets and 3D LMMs enable natural-language interaction with scenes. Popular benchmarks include ScanQA [3] and SQA3D [25]. Recent LMM-style systems include Inst3D-LLM [37], S^2-MLLM [34] , MLLM-For3D [15], SD-VLM [8], and Reason3D [16]. However, most systems answer in text only, which makes it hard to verify whether the model is referring to the correct region. In addition, metric reasoning is often treated as a side task and is not standardized across benchmarks.

Metric spatial reasoning. Several works study size, distance, and spatial relations in either images or 3D data. N3D-VLM [31] and SD-VLM [8] move toward metric reasoning in 3D settings, but metric evaluation is not consistently paired with explicit point-level grounding of the referred region. SpatialVLM [6] facilitates basic spatial query responses from single images, and SpatialRGPT [9] advances this by incorporating 3D scene-graph-derived regional awareness. Nevertheless, both frameworks remain constrained by a reliance on 2D-centric training data or external depth plugins, ultimately failing to achieve a native, metric-scale 3D understanding of physical distances and object sizes. These efforts highlight the need for numeric supervision, but they do not provide a unified setting that requires grounding a referred region and returning metric quantities for that region.

On the other hand, our proposed Ground3D-LMM method combines point-level grounding and metric reasoning in a single conversational interface and emphasizes open-vocabulary part understanding. We also provide a task and dataset that standardize evaluation for grounding and measurement.

3 Ground3D Dataset

Task Definition. We define *3D Grounded Measurement* as follows: given a 3D scene (point cloud, optionally with an aligned RGB view) and a natural-language query, the model must (i) identify the referred object or part as a point-level 3D mask, and (ii) answer with the requested physical quantity in consistent units (e.g., dimensions, thickness, clearance, or inter-object distance). Dataset statistics and the annotation pipeline are summarized in Figure 2.



Source	Level	Train	Test	Total
ScanNet (*1,433 Train / 80 Test Scenes*)				
	Object	621,110	34,650	655,760
	Part	871,612	38,293	909,905
	Multi-turn	62,064	–	62,064
ScanNet++ (*905 Train / 48 Test Scenes*)				
	Object	320,946	18,450	339,396
	Part	485,387	13,012	498,399
	Multi-turn	9,783	–	9,783
Total QA	–	**2,370,902**	**104,405**	**2,475,307**

Fig. 2: Ground3D statistics and annotation pipeline. Starting from ScanNet [12] and ScanNet++ [36] RGB-D sequences, we sample representative frames, detect object/part candidates, generate and refine 2D masks, lift them to point-level 3D supervision via depth back-projection, and compute metric attributes and spatial relations. We then synthesize grounded QA at object/part level, scene-relational level, and multi-turn dialogue, followed by automatic filtering and manual verification on an evaluation subset.

3.1 Dataset Subtasks

Ground3D dataset is organized as a set of complementary subtasks that cover grounded conversation and metric reasoning at both object and part granularity: **3D metric estimation:** report object/part dimensions derived from the referred region; **3D grounded measurement:** jointly return a mask and the requested measurement for the same referent; **3D distance queries:** compute distances/clearances between two grounded regions; **3D depth relations:** reason about closer/farther and front/behind based on 3D geometry; **3D size comparison:** compare sizes across objects/parts using consistent units; and **Spatial relations:** predict left/right/above/below relations grounded to explicit regions.

We also include subtasks that reflect interactive usage, where the model must handle absence, affordances, and context carried across turns: **Existence verification:** confirm the presence or absence of a queried object, including plausible but non-existent queries to reduce false positives; **Functional grounding:** identify parts that support an action or affordance (e.g., "the graspable handle") and ground them; and **Multi-turn conversation:** maintain context across turns while alternating between grounded and purely textual responses.

3.2 Dataset Construction

We use around 2.5K scenes from ScanNet and ScanNet++ datasets to build Ground3D dataset and benchmark. Each scene provides calibrated RGB-D frames

with camera intrinsics and poses, enabling us to lift 2D segmentations into 3D and compute metric quantities in real units. An overview of the pipeline is shown in Figure 2.

Stage 1: Frame sampling and 2D object/part grounding. For each scene, we select a fixed number of sharp, well-spread frames to cover diverse viewpoints. We prompt Qwen3-VL-30B with a structured instruction to propose visible objects and semantically meaningful parts, returning object/part names and bounding boxes in JSON format. The prompt follows a hierarchical object→part strategy and allows open-vocabulary candidates to increase part diversity beyond the native dataset labels.

Stage 2: 2D mask generation and refinement. Given the predicted boxes, we generate instance masks with SAM2 [28]. We refine masks in two ways. When a prediction can be aligned to ScanNet/ScanNet++ instances, we project masked pixels into 3D using depth and pose and keep pixels consistent with the dominant 3D instance (with name-to-label matching for the predicted category). For candidates that do not correspond to a known label, we apply a geometry-only refinement that erodes uncertain boundaries (distance-transform based), retaining a stable core region and reducing boundary noise.

Stage 3: 3D back-projection and mask cleanup. We back-project depth into a camera-frame point cloud and lift each refined 2D mask to an object/part-specific 3D point subset. To suppress depth noise and accidental leakage, we remove 3D outliers using per-axis percentiles (e.g., 5%-95%), producing clean point-level masks used for supervision and evaluation.

Stage 4: Metric and relational context generation. From each grounded point set, we compute per-instance geometric attributes including camera-frame centroid, oriented bounding box (OBB) dimensions (length, width, thickness), and camera distance. We additionally compute scene-level relations used by relational subtasks: pairwise distances/clearances (5th-percentile closest-point distance), left/right ordering (camera-x), closer/farther (camera-z), and global size ranking via OBB volume. All metrics are defined over the visible 3D geometry from the current viewport rather than the full object mesh, matching what the model perceives and reflecting embodied and robotics settings.

Stage 5: Grounded Q/A synthesis (three levels). We synthesize question–answer pairs with Qwen3-VL-8B conditioned on the computed attributes. **Level 1** generates single-object/part Q/A covering metric estimation, grounded measurement, and functional grounding. **Level 2** samples ordered object/part pairs to generate distance, depth, size, spatial, and existence queries conditioned on pairwise relations. **Level 3** assembles three-turn grounded dialogues for selected object–part pairs, with motives such as reachability, comparison, identification, organization, and preference. Across all levels, grounded spans are marked with `<p>... </p><SEG>` to align answer text with the corresponding point mask(s).

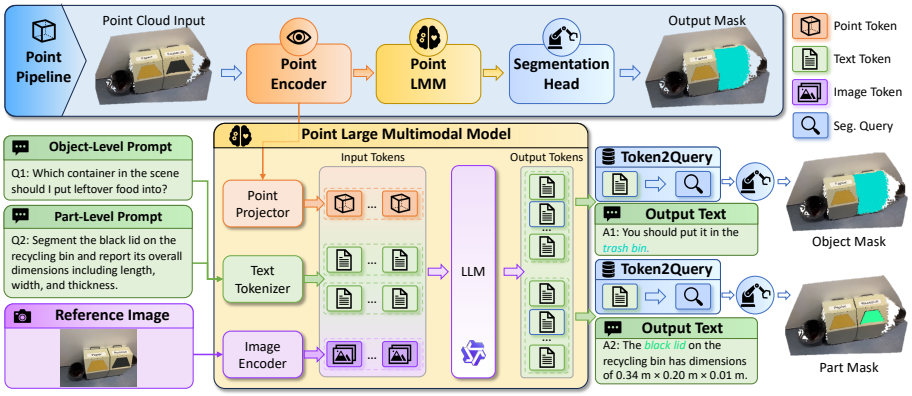


Fig. 3: Ground3D-LMM overview. Given a colored point cloud and an optional RGB image, the model answers in text and outputs point-level masks when `<SEG>` is triggered.

Verification and filtering. We apply rule-based cleanup to remove malformed outputs (e.g., broken `<p>... </p><SEG>` tags, missing numbers/units, or inconsistent metrics), prune noisy segmentations, and discard corrupted depth/metric entries. A secondary VLM-based verifier (Qwen3-VL-30B) checks (a) phrase–mask consistency, (b) metric plausibility against the image evidence, and (c) answer contradictions, dropping any item that fails. Finally, we manually inspect all 128 evaluation scenes for mask quality and metric correctness (∼26% removed); a further ∼3% is removed by VLM filtering, yielding 104,405 retained high-quality samples. The resulting dataset includes object-level, part-level, scene-level, and multi-turn grounded Q/A pairs together with aligned 3D masks and metric annotations.

4 Method

Problem Definition. We represent a colored point cloud of L points as $X \in \mathbb{R}^{L\times 6}$, where each point has coordinates and color $[x, y, z, r, g, b]$. Given a natural-language query q and an optional RGB image I our goal is to produce: (i) a textual answer y, and (ii) when the query refers to a specific region, a point-level mask $m \in \{0, 1\}^N$ that segments the referred object or part. For metric queries, the answer also includes numeric values (e.g., length, width, height, thickness, distances) in physical units.

4.1 Ground3D-LMM Overview

The overall framework of our Ground3D-LMM is as shown in Figure 3. Ground3D-LMM couples a 3D point encoder with a large multimodal language model. The point encoder produces a set of point or superpoint features. The LMM performs instruction following and generates the response text. When the output

contains a special grounding trigger token (<SEG>), we run a segmentation head that predicts a 3D mask aligned with the triggered phrase. Moreover, the LLM computes metric quantities for the predicted region (e.g., oriented bounding box dimensions, distances between regions), and these values are verbalized in the final answer.

4.2 Point Encoder

To balance computational efficiency with the retention of fine-grained spatial details, we employ a voxelization-based preprocessing strategy [10]. Specifically, the input point cloud consisting of L points is downsampled into M superpoints ($M < L$), which serves to aggregate local geometric structures while filtering out redundant noise. These superpoints are subsequently processed by a sparse 3D U-Net to extract hierarchical point-wise features. The resulting representation, $\mathcal{F}_p \in \mathbb{R}^{M \times d}$ (where d denotes the embedding dimensionality), captures both local geometry and global scene context, providing a robust foundation for subsequent multimodal alignment.

4.3 Point Large Multimodal Model (Point LMM)

The core of our framework is a Large Multimodal Model (LMM) capable of reasoning across 3D, 2D, and textual modalities.

Multimodal Alignment. To bridge the modality gap, we utilize a linear projector P to map the 3D point features into the LLM's joint latent space:

$$\mathcal{T}_p = P(\mathcal{F}_p). \tag{1}$$

These point tokens $\mathcal{T}_p$ encapsulate essential spatial priors, enabling the LLM to perceive the 3D environment. To further augment the model's grounding capability, we incorporate a camera-view reference image. The extracted image tokens $\mathcal{T}_{img}$ provide complementary visual cues (e.g., relative spatial position in the instruction) that are often ambiguous in raw point clouds, thereby facilitating more precise spatial reasoning.

Instruction Formulation. We follow a prompt-based learning paradigm where the model is tasked with answering scene-related queries regarding objects or their constituent parts. Specifically, we reserve M special tokens, denoted as <|point|>, within the input instruction. These placeholders are dynamically replaced by the embedded point tokens $\mathcal{T}_p$. To enable task-specific output, we adopt the segmentation-aware signaling mechanism where target objects or parts are wrapped within specific tags (e.g., object_name <SEG>).

The final response generation is formulated as:

$$\mathcal{Y}_{out} = \text{LMM}(\mathcal{T}_{text}, \mathcal{T}_{img}, \mathcal{T}_p), \tag{2}$$

where $\mathcal{T}_{text}$ represents the tokenized user instruction as below.

Prompt Example

```
You are an AI visual assistant analyzing a 3D point-cloud scene. Your task is to provide
spatial analysis based on the point-cloud data and the given image. Please include in
your response the objects or object parts annotated with <p>object name</p><SEG> or
<p>part name</p><SEG>.
The point-cloud data is <|point|>...<|point|>.
The question is: Which container in the scene should I put leftover food into?
```

4.4 Segmentation Head

Upon generating the textual response, we isolate the latent representations associated within the segmentation tokens, which is denoted as $\mathbf{t}$. We employ a query embedding function Q to transform these states into learnable segmentation queries $\mathbf{q}$. These queries interact with the dense point features $\mathcal{F}_p$ through a sequence of attention-based refinements:

$$\begin{aligned} \mathbf{q}^{(0)} &= Q(\mathbf{t}), & \mathbf{q}^{(1)} &= \text{Cross-Attn}(\mathbf{q}^{(0)}, \mathcal{F}_p), \\ \mathbf{q}^{(2)} &= \text{Self-Attn}(\mathbf{q}^{(1)}), & \mathbf{q}^{(f)} &= \text{FFN}(\mathbf{q}^{(2)}). \end{aligned} \tag{3}$$

Finally, a lightweight mask decoder computes the similarity between the refined query $q^{(f)}$ and the point features $\mathcal{F}_p$ to produce the final binary mask:

$$\mathcal{M} = \text{Decoder}(q^{(f)}, \mathcal{F}_p). \tag{4}$$

4.5 Training Objectives

The Ground3D-LMM is trained end-to-end using a multi-task objective function that encompasses both spatial grounding and linguistic generation. The total loss $\mathcal{L}_{total}$ is defined as a weighted sum of the segmentation loss $\mathcal{L}_{seg}$ and the language modeling loss $\mathcal{L}_{text}$:

$$\mathcal{L}_{total} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{text}\mathcal{L}_{text} \tag{5}$$

where λ_{seg} and λ_{text} are hyperparameters balancing the two objectives.

Segmentation Loss. To ensure precise mask prediction for both objects and parts, we utilize a combination of Binary Cross-Entropy (BCE) loss and Dice loss [26] for the segmentation head. While the BCE loss enforces point-wise classification accuracy, the Dice loss is employed to mitigate the class imbalance issue inherent in sparse point clouds:

$$\mathcal{L}_{seg} = \mathcal{L}_{BCE}(\mathcal{M}, \mathcal{M}_{gt}) + \mathcal{L}_{Dice}(\mathcal{M}, \mathcal{M}_{gt}) \tag{6}$$

where $\mathcal{M}$ and $\mathcal{M}_{gt}$ denote the predicted mask and the corresponding ground-truth label, respectively.

Language Modeling Loss. For the textual response generation, we adopt the standard auto-regressive cross-entropy loss. The model is optimized to predict the next token in the sequence given the multimodal context:

$$\mathcal{L}_{text} = -\sum_{i=1}^{N} \log P(y_i \mid y_{<i}, \mathcal{T}_{text}, \mathcal{T}_{img}, \mathcal{T}_p) \tag{7}$$

where y_i represents the i-th token in the target output sequence $\mathcal{Y}_{gt}$ of length N. This loss ensures that the LLM maintains its reasoning and conversational capabilities while learning to trigger the segmentation task via special tokens.

5 Experiments

5.1 Evaluation Metrics

Segmentation. We evaluate point grounding using mean IoU (mIoU) between predicted and ground-truth point masks.

Metric answers. For queries requiring numerical responses, we extract predicted scalar values using the LLM and normalize units (e.g., cm, m) to a common scale. We report Absolute Percentage Error (APE), defined as $\text{APE} = \left|\frac{\hat{s}-s}{s}\right| \times 100\%$, as well as the δ success rate, where $\delta = \max\left(\frac{\hat{s}}{s}, \frac{s}{\hat{s}}\right)$ and a prediction is considered correct if $\delta \leq \tau$ (with $\tau = 1.25$). Here, s denotes the ground-truth physical quantity and $\hat{s}$ the predicted value.

Grounded Measurement. To jointly assess grounding and metric accuracy, we introduce the *Grounded-Measurement Success Rate* (GM-δ). A prediction is counted as successful only when its mask IoU with the ground truth exceeds 0.3 *and* its metric prediction satisfies $\delta \leq 1.25$. This unified metric integrates segmentation quality and metric correctness into a single score; results are reported in the supplementary material.

Text quality. To compare text outputs across models, we report a rubric-based evaluation over two criteria: *Hallucination* (penalizing unsupported claims about the scene), and *Completeness* (whether the response addresses all requested quantities). Each criterion is scored on a 0–10 scale. Scores are averaged over the Ground3D evaluation split; the same judging protocol is applied to all methods.

5.2 Experimental Setup

We train Ground3D-LMM on our dataset and evaluate it on both object-level and part-level tasks. We further evaluate the model on the Reason3D dataset and ScanRefer with additional fine-tuning. For the Ground3D dataset, we adopt our predefined data splits for all experiments. For other datasets, we follow the corresponding splits for training and finetuning. We report results for both 3D-only (point cloud) and 3D+2D (point cloud + a reference RGB view) settings.

Table 2: Object segmentation results (mIoU) on Ground3D-ScanNet and Ground3D-ScanNet++. Our Ground3D-LMM achieves superior performance across both 3D-only and multimodal (3D+2D) settings.

Method	Modality	Overall	Dist.	Func. Obj.	Gnd. Mea.	Depth Rel.	Spa. Rel.	Size Comp.	Metric Est.
Ground3D-ScanNet									
Image baseline	2D	32.52	31.46	36.67	37.51	32.08	28.11	25.66	36.18
Reason3D [16]	3D	9.31	–	6.09	12.07	–	–	–	9.76
UniSeg3D [35]	3D	22.97	23.64	24.00	23.87	22.56	21.71	21.13	23.87
Ground3D-LMM	3D	37.40	37.79	37.64	39.10	35.79	37.42	34.89	39.14
Ground3D-LMM	3D+2D	**42.22**	**41.37**	**43.73**	**44.62**	**42.59**	**40.42**	**38.56**	**44.28**
Ground3D-ScanNet++									
Image baseline	2D	27.80	27.97	23.56	30.08	27.70	28.21	28.49	28.60
Reason3D [16]	3D	2.89	–	1.34	3.95	–	–	–	3.39
UniSeg3D [35]	3D	10.84	11.84	10.86	10.83	11.16	10.95	9.39	10.82
Ground3D-LMM	3D	24.56	24.70	18.56	26.04	24.11	26.86	25.85	25.79
Ground3D-LMM	3D+2D	**30.42**	**30.97**	**20.27**	**33.08**	**31.60**	**34.55**	**32.44**	**30.01**

Implementation Details. For point encoder, we adopt a sparse convolutional backbone [11] integrated with superpoint pooling [30], which is initialized with pre-trained weights from SSTNet [22]. For the LMM, we leverage the Qwen3-VL-4B-Instruct [4]. Both the Point Projector and the query embedding function are implemented as Multi-Layer Perceptrons (MLPs). Our framework is optimized using the AdamW [24] optimizer with a weight decay of 0.05. To accommodate the varying convergence rates of different components, we apply a stratified learning rate strategy: the point encoder is fine-tuned with an initial rate of 1×10^{-6}, the LMM backbone is updated at 1×10^{-5}, while all other task-specific modules (e.g., projectors and segmentation head) are initialized at 1×10^{-4}. We employ a polynomial learning rate scheduler with a power of 0.95 and set the batch size to 1. We utilize LoRA [14] for the language model part. The loss weights are unified to $\lambda = 1.0$ without the need for extensive tuning.

5.3 Quantitative Results

Baseline Methods We compare our Ground3D-LMM with 3 main baseline: Image baseline, UniSeg3D [35] and Reason3D [16]. We construct an image-only baseline by providing the corresponding RGB frame for each question to Qwen-VL, which generates textual and metric answers with grounded segmentation phrases. The predicted phrases are localized using Grounded Sam [29] to obtain 2D masks, which are then back-projected into 3D by mapping mask pixels through the depth image using camera intrinsics. UniSeg3D [35] supports open-vocabulary 3D segmentation using text prompts as input. In contrast, Reason3D [16] performs reasoning-based segmentation, where the prompts do not explicitly specify category names but instead require the model to infer the target class.

Table 3: Part segmentation results (mIoU) on Ground3D-ScanNet and Ground3D-ScanNet++. Our method demonstrates strong generalization in fine-grained part reasoning, especially when integrated with 2D visual cues.

Method	Modality	Overall	Dist.	Exist.	Func. Part	Gnd. Mea.	Depth Rel.	Spa. Rel.	Size Comp.	Metric Est.
Ground3D-ScanNet										
Image baseline	2D	21.52	20.70	22.94	24.73	22.82	20.21	15.35	20.14	25.28
Reason3D [16]	3D	8.19	–	–	7.04	9.35	–	–	–	8.18
UniSeg3D [35]	3D	11.28	11.71	9.70	10.14	12.25	11.69	10.77	11.70	12.25
Ground3D-LMM	3D	30.06	30.47	32.20	27.27	30.83	30.36	28.15	30.31	30.89
Ground3D-LMM	3D+2D	**36.57**	**37.33**	**38.21**	**31.70**	**38.24**	**37.01**	**34.79**	**37.01**	**38.28**
Ground3D-ScanNet++										
Image baseline	2D	25.86	25.66	29.59	26.86	27.00	24.39	22.33	24.90	26.13
Reason3D [16]	3D	4.91	–	–	4.04	4.92	–	–	–	5.77
UniSeg3D [35]	3D	8.03	8.62	10.20	7.27	7.18	7.75	7.75	8.24	7.28
Ground3D-LMM	3D	29.80	30.69	34.07	27.02	28.77	29.81	28.93	29.62	29.49
Ground3D-LMM	3D+2D	**39.07**	**40.14**	**43.48**	**35.39**	**39.37**	**39.03**	**37.56**	**38.82**	**38.80**

Table 4: Text and metric evaluation on Ground3D-ScanNet and Ground3D-ScanNet++ evaluation subsets. Higher is better except Mean APE (lower is better).

Method	Modality	*Ground3D ScanNet Subset*				*Ground3D ScanNet++ Subset*			
		Halluc.	Comp.	Mean APE	δ Succ. Rate	Halluc.	Comp.	Mean APE	δ Succ. Rate
Image baseline	2D	8.06	7.37	166.91	25.05	8.29	7.55	120.22	24.17
SD-VLM [8]	2D	8.01	8.27	231.03	20.12	8.19	7.64	183.93	15.46
Ground3D-LMM	3D	9.30	8.38	79.85	34.78	9.30	7.97	84.15	27.49
Ground3D-LMM	3D+2D	9.16	7.99	74.03	43.55	9.05	7.39	88.20	33.93

Results on Object Segmentation We report object-level segmentation results on our Ground3D dataset in Table 2. Our Ground3D-LMM consistently demonstrates stronger performance than all baselines. Furthermore, incorporating image information brings additional performance gains, enabling Ground3D-LMM to achieve the best results across all tasks. We observe that Reason3D [16], which is designed for semantic segmentation, consistently underperforms on our dataset. This is because our tasks require precise identification of a specific individual object. Moreover, some tasks involve multiple objects from different classes, a setting for which Reason3D [16] is not well suited.

Results on Part Segmentation We further evaluate the methods on more fine-grained part segmentation tasks, with the results reported in Table 3. As expected, the performance of all methods drops compared to object-level segmentation, since part segmentation is inherently more challenging and requires finer-grained 3D perception. Nevertheless, our method still achieves the best performance on this task. Even when using only point cloud input, our approach outperforms all baselines across all tasks.

Table 5: Segmentation results on Reason3D. The evaluation metrics include accuracy at IoU thresholds 0.25, 0.50, and mean IoU.

Methods	Venue	Modality	Acc@0.25	Acc@0.50	mIoU
3D-STMN [32]	AAAI'24	3D	25.43	17.78	18.23
Intent3D [19]	ICLR'25	3D	20.57	19.46	-
Reason3D [16]	3DV'25	3D	43.21	32.10	31.20
Ground3D-LMM (Ours)	-	3D	**50.32**	**37.01**	**36.35**
MLLM-For3D [15]+ LISA-7B	NeurIPS'25	3D+2D	45.53	39.54	31.94
MLLM-For3D [15]+ VideoLISA	NeurIPS'25	3D+2D	48.45	41.02	39.82
Ground3D-LMM (Ours)	-	3D+2D	**57.79**	**41.23**	**41.29**

Results on Text Response In addition to mask quality, we evaluate response quality and metric accuracy in Table 4. Across Ground3D-ScanNet and Ground3D-ScanNet++ evaluation subsets, Ground3D-LMM consistently outperforms the image-only baseline and SD-VLM [8], with further gains when combining 3D and 2D cues. Similar improvements are obeserved on both subsets, demonstrating strong cross-dataset generalisation. In text and metric evaluation, our model achieves lower Mean APE and higher δ success rate, indicating more accurate quantitative reasoning and improved geometric grounding compared to purely image-based approaches.

Results on Reason3D Table 5 compares Ground3D-LMM to Reason3D [16] and MLLM-For3D [15] on ScanNet scenes following the Reason3D [16] protocol. Using only point cloud input, our method achieves the best performance, surpassing Reason3D [16] by a large margin (5.15% on mIoU). When further incorporating image information, Ground3D-LMM consistently outperforms all prior methods and delivers the strongest overall results. Notably, even without image inputs and using only a 4B backbone, Ground3D-LMM still significantly outperforms MLLM-For3D equipped with a 7B model and image inputs, demonstrating the effectiveness and efficiency of our design.

Results on ScanRefer We further evaluate open-vocabulary grounding on ScanRefer [7], which poses a more challenging setting than Reason3D [16]. Unlike category-level semantic segmentation, ScanRefer demands instance-level grounding, where the model must localize and segment the unique target object described by a referring expression. This requires not only semantic understanding but also fine-grained discrimination among multiple objects of the same category. With only 3D input, our method already surpasses existing approaches by nearly 20%, demonstrating strong instance-level reasoning capability from pure point cloud features. When image inputs are incorporated, performance is further improved, exceeding the second-best method by 10%. These results indicate that our approach is effective not only for question-driven semantic segmentation, but also for more challenging single-instance open-vocabulary grounding.

Table 6: Evaluation results on ScanRefer dataset. The evaluation metrics include accuracy at IoU thresholds 0.25, 0.50, and mean IoU.

Methods	Venue	Modality	Acc@0.25 ↑	Acc@0.50 ↑	mIoU ↑
ScanRefer [7]	ECCV'20	3D	10.51	6.20	-
TGNN [17]	AAAI'21	3D	11.64	9.51	8.13
InstanceRefer [38]	ICCV'21	3D	13.92	11.47	-
BUTD-DETR [18]	ECCV'22	3D	11.99	8.95	-
M3DRef-CLIP [39]	ICCV'23	3D	18.3	14.8	10.29
EDA [33]	CVPR'23	3D	26.50	21.20	-
Reason3D [16]	3DV'25	3D	17.64	13.11	13.05
IntentNet [19]	ICLR'25	3D	28.12	22.63	18.92
UniSeg3D [35]	NeurIPS'24	3D	-	-	29.10
Ground3D-LMM (Ours)	-	3D	**55.73**	**32.71**	**38.72**
MLLM-For3D [15]+ LISA-7B	NeurIPS'25	3D+2D	31.88	29.90	28.10
MLLM-For3D [15]+VideoLISA	NeurIPS'25	3D+2D	33.12	31.21	30.45
Ground3D-LMM (Ours)	-	3D+2D	**59.98**	**36.37**	**41.30**

Table 7: Ablation study on the object-level task (Ground3D-ScanNet). We ablate *(a)* dataset scale, *(b)* joint mask + metric supervision, and *(c)* segmentation supervision quality. Higher is better except APE (lower is better).

Setting	mIoU↑	APE↓	δ ↑	Halluc.↑	Comp.↑
Ground3D-LMM (Ours, 3D+2D)	42.22	76.07	46.34	9.16	7.77
(a) Ours w. 25% data	36.93	99.23	40.26	8.86	7.03
(a) Ours w. 50% data	39.64	81.89	40.84	8.93	7.26
(b) Ours w. text-only	–	77.41	45.62	9.02	7.35
(c) Ours w. 25% noise mask	41.43	77.97	44.09	8.91	7.32
(c) Ours w. 50% noise mask	41.06	79.60	43.84	9.01	7.45

5.4 Ablation Study

We ablate three design choices on Ground3D-ScanNet object-level results (Table 7): *(a)* reducing training data to 25%/50% degrades all metrics, confirming the full dataset is necessary; *(b)* text-only supervision eliminates mask prediction and weakens metric quality, showing joint mask and metric supervision are mutually beneficial; *(c)* randomly zeroing up to 50% of ground-truth masks causes only modest mIoU drops, indicating robustness to noisy segmentation supervision.

5.5 Qualitative Results

In Fig. 4, we present qualitative results of Ground3D-LMM across six sub-tasks. Our model performs well on tasks involving relative position, scale comparison, and metric estimation. For example, in the Scale Comparison and Grounded Dimension task, the prediction error is within 0.05 m. Moreover, from functional

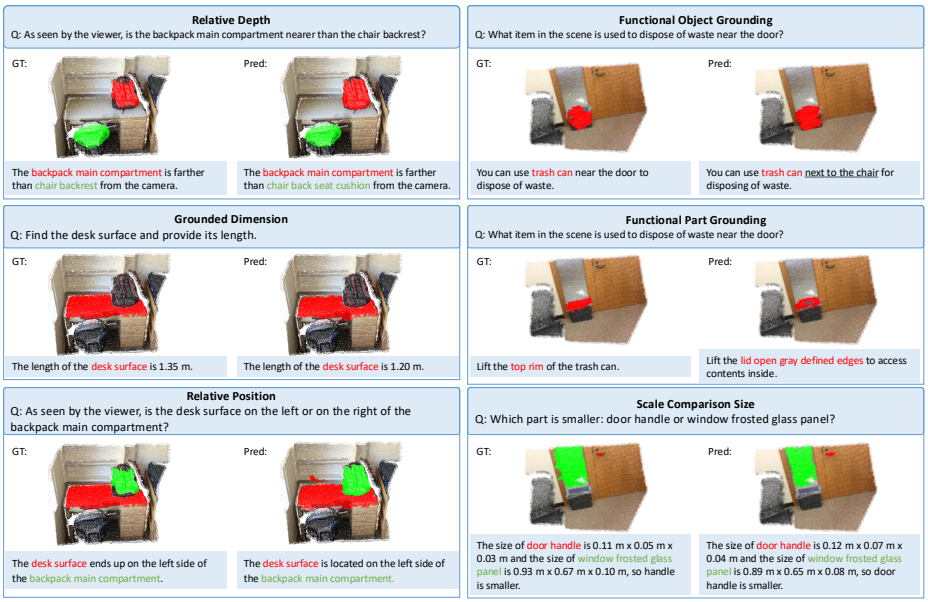


Fig. 4: Qualitative results of Ground3D-LMM across six representative object- and part-level sub-tasks. (Zoom in for better visualization.)

object grounding to functional part grounding, the model accurately localizes and describes both objects and their corresponding parts within the same scene. The consistent performance at both object and part levels highlights the effectiveness of our method and the validity of the proposed dataset.

6 Conclusion

We introduced Ground3D-LMM, a unified point-cloud multimodal model that produces verifiable 3D conversational outputs by pairing natural-language responses with point-level masks, and supports metric measurements in consistent real-world units. We also proposed the 3D Grounded Measurement task and the Ground3D dataset, which jointly benchmark object- and part-level grounding, multi-object spatial relations, and multi-turn dialogue. Experiments on Ground3D, Reason3D, and ScanRefer demonstrate strong gains over open-vocabulary segmentation and 3D reasoning baselines, and highlight the benefit of incorporating an RGB view when available.

Acknowledgement

The computations were enabled by resources provided by LUMI hosted by CSC (Finland) and LUMI consortium, and by Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the NSC.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In: European conference on computer vision. pp. 422–440. Springer (2020)
2. Ahmed, M., Fei, J., Ding, J., Bakr, E.M., Elhoseiny, M.: Kestrel: 3D multimodal llm for part-aware grounded description. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8973–8983 (2025)
3. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3D question answering for spatial scene understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19129–19139 (2022)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Boudjoghra, M.E.A., Dai, A., Lahoud, J., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.S.: OPEN-YOLO 3D: Towards fast and accurate open-vocabulary 3D instance segmentation. In: 13th International Conference on Learning Representations, ICLR 2025. pp. 26618–26631. International Conference on Learning Representations, ICLR (2025)
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
7. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3D object localization in rgb-d scans using natural language. In: ECCV. pp. 202–221. Springer (2020)
8. Chen, P., Lou, Y., Cao, S., Guo, J., Fan, L., Wu, Y., Yang, L., Ma, L., Ye, J.: SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. arXiv preprint arXiv:2509.17664 (2025)
9. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: SpatialRGPT: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
10. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
11. Contributors, S.: Spconv: Spatially sparse convolution library. `https://github.com/traveller59/spconv`, last accessed 2026/06/26 (2022)
12. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
13. He, S., Ding, H., Jiang, X., Wen, B.: Segpoint: Segment any point cloud via large language model. In: ECCV. pp. 349–367. Springer (2024)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
15. Huang, J., et al.: MLLM-For3D: Adapting multimodal large language model for 3D reasoning segmentation. In: NeurIPS (2025)
16. Huang, K.C., Li, X., Qi, L., Yan, S., Yang, M.H.: Reason3D: Searching and reasoning 3D segmentation via large language model. arXiv preprint arXiv:2405.17427 (2024)
17. Huang, P.H., Lee, H.H., Chen, H.T., Liu, T.L.: Text-guided graph neural networks for referring 3D instance segmentation. AAAI **35**(2), 1610–1618 (May 2021)

18. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: Bottom up top down detection transformers for language grounding in images and point clouds. arXiv preprint arXiv:2112.08879 (2021)
19. Kang, W., Qu, M., Kini, J., Wei, Y., Shah, M., Yan, Y.: Intent3D: 3D object detection in rgb-d scans based on human intention. arXiv preprint arXiv:2405.18295 (2024)
20. Kareem, A., Lahoud, J., Cholakkal, H.: Paris3D: Reasoning-based 3D part segmentation using large multimodal model. In: European Conference on Computer Vision. pp. 466–482. Springer (2024)
21. Li, R., Li, S., Kong, L., Yang, X., Liang, J.: Seeground: See and ground for zero-shot open-vocabulary 3D visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3707–3717 (2025)
22. Liang, Z., Li, Z., Xu, S., Tan, M., Jia, K.: Instance segmentation in 3D scenes using semantic superpoint tree networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2783–2792 (2021)
23. Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., Su, H.: 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. pp. 21736–21746 (2023)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
25. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: SQA3D: Situated question answering in 3D scenes. In: International Conference on Learning Representations (2023)
26. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
27. Nguyen, P., Ngo, T.D., Kalogerakis, E., Gan, C., Tran, A., Pham, C., Nguyen, K.: Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4018–4028 (2024)
28. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
29. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded SAM: Assembling open-world models for diverse visual tasks (2024). https://doi.org/10.48550/arXiv.2401.14159
30. Robert, D., Raguet, H., Landrieu, L.: Efficient 3D semantic segmentation with superpoint transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
31. Wang, Y., Ke, L., Zhang, B., Qu, T., Yu, H., Huang, Z., Yu, M., Xu, D., Yu, D.: N3D-VLM: Native 3D grounding enables accurate spatial reasoning in vision-language models. arXiv preprint arXiv:2512.16561 (2025)
32. Wu, C., Ma, Y., Chen, Q., Wang, H., Luo, G., Ji, J., Sun, X.: 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. vol. 38, pp. 5940–5948 (2024)
33. Wu, Y., Cheng, X., Zhang, R., Cheng, Z., Zhang, J.: Eda: Explicit text-decoupling and dense alignment for 3D visual grounding. In: CVPR. pp. 19231–19242 (2023)

34. Xu, B., Zhu, S., Jin, Z., Li, J., Wang, H.: S^2-MLLM: Boosting spatial reasoning capability of mllms for 3D visual grounding with structural guidance. arXiv preprint arXiv:2512.01223 (2025)
35. Xu, W., Shi, C., Tu, S., Zhou, X., Liang, D., Bai, X.: A unified framework for 3D scene understanding. Advances in Neural Information Processing Systems **37**, 59468–59490 (2024)
36. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3D indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
37. Yu, H., Li, W., Wang, S., Chen, J., Zhu, J.: Inst3d-lmm: Instance-aware 3D scene understanding with multi-modal instruction tuning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14147–14157 (2025)
38. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1791–1800 (October 2021)
39. Zhang, Y., Gong, Z., Chang, A.X.: Multi3drefer: Grounding text description to multiple 3D objects. In: ICCV. pp. 15225–15236 (2023)