This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Scenes as Tokens: Multi-Scale Normal Distributions Transform Tokenizer for General 3D Vision–Language Understanding

Yutao Tang*[1], Cheng Zhao[2], Gaurav Mittal[2], Rohith Kukkala[2],
Rama Chellappa[1], Cheng Peng[1], Mei Chen[2]
[1]Johns Hopkins University, [2]Microsoft
{ytang67, rchella4, cpeng26}@jhu.edu,
{cheng.zhao, gaurav.mittal, rohith.kukkala, mei.chen}@microsoft.com

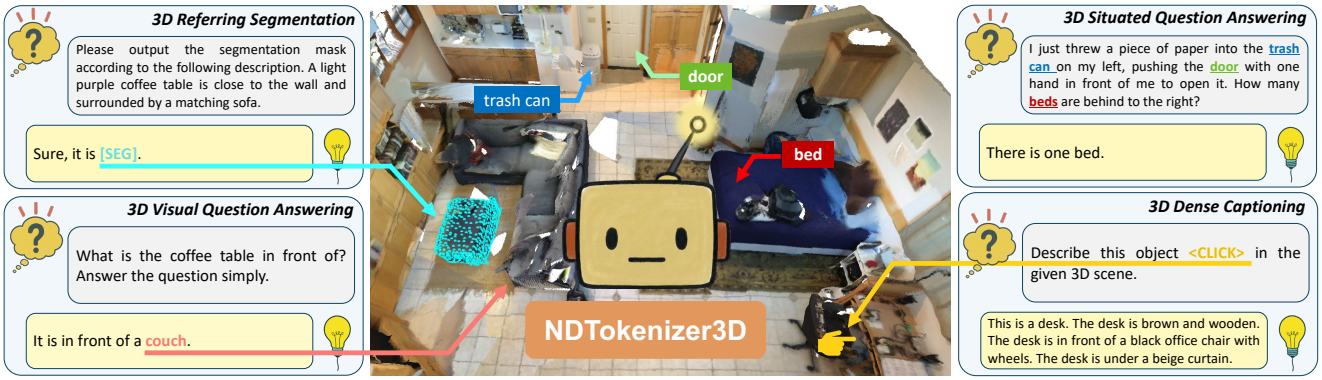

Figure 1. We introduce NDTokenizer3D, a generalist 3D VLM that bridges language-level reasoning with spatial understanding. By tokenizing complex 3D scenes into information-rich representations, NDTokenizer3D enables diverse tasks such as 3D Visual Question Answering, Dense captioning, and Referring Segmentation within a unified and interactive framework.

Abstract

Recent advances in 3D vision–language models (VLMs) highlight a strong potential for 3D scene understanding and reasoning. However, effectively tokenizing 3D scenes into holistic scene tokens, and leveraging these tokens across diverse 3D understanding tasks, remain highly challenging. We present NDTokenizer3D, a generalist 3D VLM that performs a wide range of 3D scene understanding tasks while naturally supporting human interactions, thereby bridging language-level reasoning with 3D spatial understanding. The core of our approach is a novel three-stage scene tokenization pipeline built upon a Multi-Scale Normal Distributions Transform (NDT) representation, paired with a Multi-Scale NDT Decoder (MSDec). Specifically, NDTokenizer3D first constructs a multi-scale NDT representation from raw high-resolution point clouds, preserving both global context and fine-grained geometric details. Next, the MSDec progressively fuses cross-scale NDT features, producing holistic scene tokens consumable by LLM endpoints. Beyond tokenization, MSDec is repurposed as a general interface for human-interactive prompting (points, boxes, masks) and segmentation-mask decoding, unifying diverse 3D scene understanding tasks within a single architecture. With this compact and unified design, NDTokenizer3D offers a fine-grained, general-purpose 3D VLM, achieving remarkable improvements in 3D Referring Segmentation, 3D Visual Question Answering, and 3D Dense Captioning.

1. Introduction

Extending vision–language modeling from 2D to 3D opens the door to a range of impactful applications, such as autonomous driving, embodied AI , and 3D AR/VR, in which robust reasoning, fine-grained 3D scene understanding, and seamless human–agent interaction are essential. Unlike conventional 2D VLMs, 3D VLMs require an effective 3D scene tokenization method to process 3D data. The core challenge lies in compressing massive high-resolution point clouds into bounded-length token sequences for an LLM endpoint while minimizing information loss. Previously re-

*This work was done during Yutao's internship at Microsoft.
†Code: https://github.com/snldmt/NDTokenizer3D

ported point-cloud–based tokenization pipelines [8, 16, 64] process raw point clouds through downsampling, sacrificing fine-grained 3D geometric details and lacking an effective mechanism to capture the abstract global structure. Moreover, achieving comprehensive 3D scene understanding requires modeling object-environment and inter-object relationships across multiple scales, as objects naturally appear at varying spatial resolutions. Consequently, it is crucial to capture both local details and broad contextual cues for holistic understanding. However, most current approaches either treat an entire 3D scene as a single-scale entity [8, 15] or represent a scene using a series of object instances [10, 21], thereby neglecting the intricate multi-scale relationships that underpin realistic spatial reasoning.

To address these issues, we adopt a compact multi-scale Normal Distributions Transform (NDT) [49] representation that operates directly on high-resolution point clouds without downsampling. Originally developed for Simultaneous Localization and Mapping (SLAM), NDT partitions a 3D point cloud into a uniform grid and models each cell's local surface as a Gaussian distribution. This grid-based formulation naturally supports multiple resolutions: fine-resolution cells capture detailed local geometry, while coarse-resolution cells encode global scene structure. Unlike naive downsampling, the fine-resolution NDT retains information from raw points via Gaussian statistics (mean and covariance) inside the NDT cell, preserving local geometric details maximally. In parallel, the coarse-resolution NDT aggregates over larger spatial regions, providing an abstract representation that captures region-level global context information. To effectively integrate these multi-scale representations, we further design a Multi-Scale NDT Decoder (MSDec) that progressively fuses features from coarse to fine scales, producing holistic scene tokens consumable by LLM endpoints.

Another challenge of 3D VLM lies in the limited support for human interaction [23, 64] and the task-specific nature of existing methods [17, 25, 39, 47], which fundamentally restricts their ability to generalize across diverse scene understanding tasks. To overcome this limitation, we discard any task-tailored modules and instead deploy MSDec as a multi-purpose interface that supports both human interactive prompting and segmentation decoding within a streamlined framework. Concretely, MSDec processes user-provided visual prompts, such as points, bounding boxes, and masks, and embeds them into a guidance token, input to LLM along with the scene tokens for interactive 3D scene understanding. In addition, MSDec can also consume a special segmentation query token, from which it derives a 3D-aware segmentation feature that is decoded to produce segmentation masks.

Combining the strengths developed above, we present NDTokenizer3D, a generalist 3D VLM capable of performing a wide range of 3D scene understanding tasks while naturally supporting human interactions, bridging language-level reasoning with spatial understanding. NDTokenizer3D introduces a novel three-stage scene tokenization pipeline that represents 3D environments as information-preserving scene tokens. In stage one, NDTokenizer3D constructs a multi-scale NDT representation that captures both fine-grained local geometric details and abstract global contextual cues. In stage two, a 3D encoder extracts multi-scale features from these complementary NDT representations. In stage three, the extracted features are integrated through the MSDec, which progressively fuses information across scales to generate holistic scene tokens, enabling comprehensive reasoning over complex 3D environments. Beside tokenization, MSDec is repurposed as a general interface that enables human interactivity and segmentation mask decoding within the same framework. It can interpret user-provided visual prompts and incorporate them into the reasoning process for interactive 3D scene understanding. Additionally, the same decoder processes segmentation query to generate fine-grained 3D mask predictions, extending language reasoning naturally into spatial prediction. Our contributions are summarized as follows:

- We propose a novel three-stage scene tokenization pipeline that leverages a multi-scale NDT representation to preserve both global context and fine-grained geometry, and integrates them through a MSDec to yield information-rich scene tokens.
- We develop NDTokenizer3D as a generalist 3D VLM that supports human interactive prompting and segmentation decoding through a multi-purpose interface, unifying diverse 3D scene understanding tasks within a single architecture.
- We demonstrate that NDTokenizer3D achieves strong performance across four 3D vision–language benchmarks, showcasing its effectiveness and versatility across text generation and spatial reasoning tasks.

2. Related Work

3D Scene Understanding. Research at the intersection of vision and language has advanced a spectrum of 3D scene understanding tasks. Among them, 3D Visual Grounding [1, 5, 24, 59, 63] and Referring Segmentation [17, 23, 43, 53] focus on spatial understanding, where the goal is to localize or segment the object regions in 3D corresponding to natural language expressions. Moreover, 3D Question Answering [2, 38, 42] and Dense Captioning [7, 9, 11, 27, 29, 60] emphasize semantic and relational reasoning. These tasks require the model to understand object attributes, spatial relations, and scene context to generate textual responses grounded in the 3D environment. Early progress in this field has largely focused on specialized architectures. Many approaches are single-task

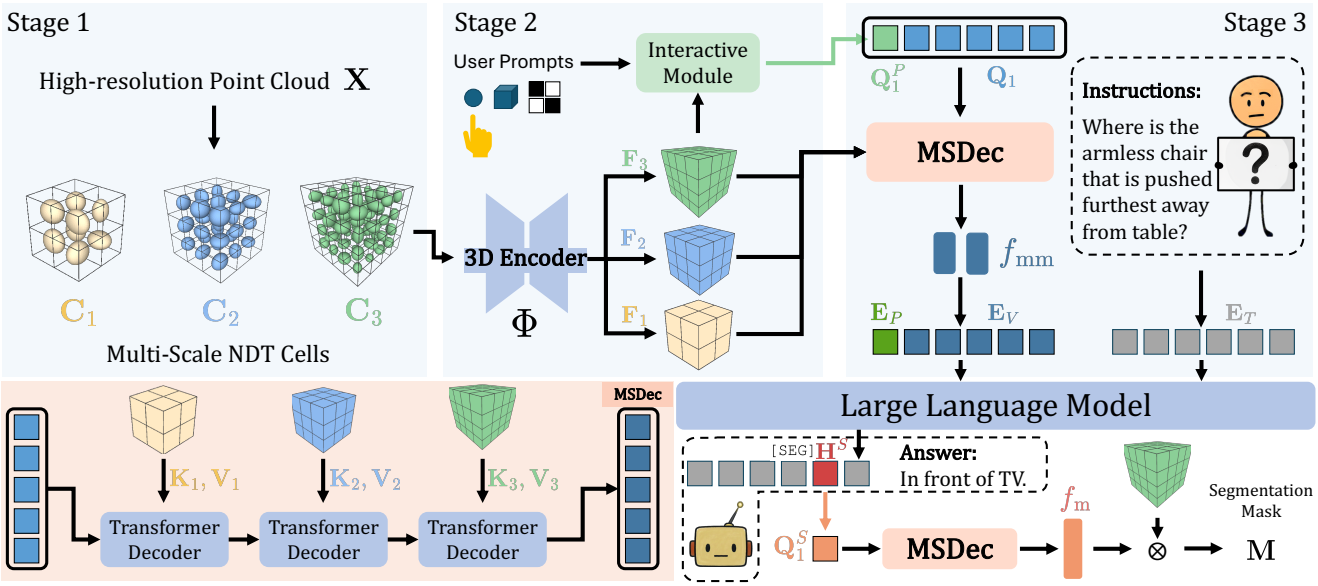

Figure 2. NDTokenizer3D is a general-purpose 3D VLM that supports a wide range of 3D understanding tasks. The model introduces a novel three-stage scene tokenization pipeline that constructs multi-scale NDT representations and aggregates them via MSDec to generate holistic scene tokens. The lower-left shows MSDec with R transformer decoder layers that integrate multi-scale NDT features, using them as `Key` and `Value`. Beyond feature integration, MSDec also acts as a unified interface for user prompting and mask decoding.

models [9, 26, 51, 55] designed for one specific objective, limiting their ability to generalize across diverse 3D understanding tasks. More recent efforts explore unified frameworks [4, 6, 12] and pretraining strategies [28, 56, 62, 65] that learn discriminative 3D features to enable shared reasoning across tasks. Nevertheless, these systems typically follow a pre-train–then–fine-tune paradigm, relying on task-specific modules or per-task finetuning after pre-training. This reliance on extra components limits their flexibility and scalability, making it difficult for them to support more natural, human-like multimodal interaction and general-purpose 3D understanding.

3D Vision Language Models. Built upon 2D VLMs, many efforts extend them into the 3D domain by injecting geometric information into LLMs for 3D understanding. A core challenge in this process lies in how 3D data are represented and aligned with language, so that the resulting features can be effectively interpreted and reasoned upon by the LLMs. Some approaches enhance 2D VLMs with limited 3D reasoning capabilities, *e.g.* by introducing bird's-eye view (BEV) image [35, 61] or decomposing scenes into object sequences [37]. While convenient, such extensions lack full 3D awareness and lead to limited accuracy. Other works [16, 20, 31] align 3D features to language through linear or MLP projections, whereas transformer-based modules [22, 58] provide richer modality interactions. More advanced methods [10, 18] employ cross-modal query mechanisms such as Q-Former [30] to facilitate 3D–language fusion. Despite these advances, tokenizing 3D scenes into information-rich and LLM-consumable representations remains an open problem. We address this challenge through a novel three-stage tokenization pipeline that produces globally informed, locally detailed 3D scene tokens for strong multimodal reasoning.

3. Methods

Figure 2 presents an overview of the NDTokenizer3D framework, illustrated with a three-scale example. It is designed as a general-purpose 3D VLM that flexibly integrates visual and textual inputs, enabling 3D vision-grounded dialogue, dense description, and referring segmentation. NDTokenizer3D begins by tokenizing 3D scenes into holistic scene tokens through a three-stage pipeline (Sec. 3.1). First, it constructs a multi-scale NDT representation from the high-resolution raw 3D point cloud as input (Sec. 3.1.1). Next, a 3D Encoder extracts multi-scale 3D features from this representation (Sec. 3.1.2). Finally, a Multi-Scale NDT Decoder (MSDec) aggregates information across scales to produce holistic 3D scene tokens (Sec. 3.1.3). In addition to feature aggregation, MSDec also serves as a unified interface that supports both user-input prompting and segmentation mask decoding (Sec. 3.2). Subsequently, we describe how the LLM utilizes all the tokens to produce task-specific outputs for various tasks in Sec. 3.3. Finally, we present the training strategy in Sec. 3.4.

3.1. 3D Scene Tokenization

3.1.1. Multi-Scale NDT

Normal Distributions Transform [49] provides a compact way to represent 3D scene data. It partitions a point cloud

into regular cells and models the local surface within each cell as a Gaussian distribution. Leveraging this grid-based formulation, we construct a multi-scale NDT representation, where the fine-scale cells preserve detailed local geometric details, avoiding substantial information loss caused by aggressive downsampling, while the coarse-scale NDT cells captures abstract global semantics. Specifically, given the high-resolution point cloud $\mathbf{X} = \{x_i\}_{i=1}^{N_p} \in \mathbb{R}^{N_p \times 3}$, each NDT cell C_r^j is characterized by its mean $\mu_r^j \in \mathbb{R}^3$ and covariance $\mathbf{\Sigma}_r^j \in \mathbb{R}^{3\times 3}$:

$$\mu_r^j = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \mathbf{\Sigma}_r^j = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_r^j)(x_i - \mu_r^j)^T, \tag{1}$$

where $\{x_i\}_{i=1}^{n}$ are the 3D points inside each cell, and $r = 1, \cdots, R$ with $r = 1$ representing the coarsest scale and $r = R$ the finest. These probability density functions implicitly interpret the structural and geometrical information of the local 3D points.

3.1.2. Feature Extraction

In addition to the mean and covariance, we also incorporate RGB color information as part of each cell's descriptor to enrich its visual representation. The cell-wise RGB values c_r^j are obtained through multi-view projection from 2D images $\mathbf{I} = \{I_k\}_{k=1}^{N_I}$ by

$$c_r^j = \frac{1}{N_I}\sum_{k=1}^{N_I} I_k(u_k, v_k),\ [u_k, v_k]^T = P(\mu_r^j|k), \tag{2}$$

where k denotes a specific view, P is the 3D-to-2D projective function, and $[u_k, v_k]^T$ are image coordinates. As a result, each cell is represented by the concatenation of its mean, covariance, and RGB values as

$$\mathbf{C}_r = \left\{C_r^j\right\}_{j=1}^{N_r} \in \mathbb{R}^{N_r \times 15},\ C_r^j = \left[\mu_r^j; \mathbf{\Sigma}_r^j; c_r^j\right]. \tag{3}$$

After obtaining NDT cells, we employ a transformer-based 3D Encoder Φ to derive multi-scale NDT features $\mathbf{F}_r = \Phi(\mathbf{C}_r) \in \mathbb{R}^{N_r \times d_f}$ where d_f is the dimension of the features. The coarse-scale features emphasize global context such as overall scene layout, environmental context, and inter-object relationships, while the fine-scale features encode detailed geometric and textural cues that describe intra-object structures and fine-grained boundaries. These encoded features serve as the basis for our multi-scale decoding framework, which integrates and refines information across different resolutions.

3.1.3. Multi-Scale NDT Decoder

To effectively integrate information across different scales, we design a transformer-based Multi-Scale NDT Decoder (MSDec), which learns a holistic and semantically rich scene representation capturing both global context and local details essential for accurate 3D scene reasoning. Specifically, the MSDec consists of R transformer decoder layers where we employ the multi-scale NDT features to serve as `Key` and `Value` inputs in each decoder layer, while a set of learnable queries serves as initial `Query`. In the first decoder layer, we initialize the query tokens using a downsampled subset of the finest-scale features by $\mathbf{Q}_1 = \mathbf{W}_1^Q(\downarrow \mathbf{F}_R)$ where $\mathbf{W}_1^Q$ is a linear projection and $\downarrow$ denotes the downsample operator. Each MSDec layer then updates the queries through a cross-attention layer with multi-scale NDT features, followed by a self-attention layer and a feed-forward network (FFN) layer as

$$\begin{aligned} \tilde{\mathbf{Q}}_r &= \mathrm{CrossAttn}(\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r) \\ \hat{\mathbf{Q}}_r &= \mathrm{SelfAttn}(\tilde{\mathbf{Q}}_r),\ \ \mathbf{Q}_{r+1} = \mathrm{FFN}(\hat{\mathbf{Q}}_r), \end{aligned} \tag{4}$$

where $\mathbf{K}_r = \mathbf{W}_r^K \mathbf{F}_r, \mathbf{V}_r = \mathbf{W}_r^V \mathbf{F}_r$. The final output $\mathbf{Q}_R$ encapsulates the holistic 3D scene representation, integrating information from all scales. This representation is then projected via a multimodal alignment head f_{mm} into the LLM input space, yielding the 3D scene tokens $\mathbf{E}_V = f_{\text{mm}}(\mathbf{Q}_R)$, serving as visual context. Through this hierarchical decoding process, MSDec effectively fuses global semantics with fine-grained details, tokenizing a 3D scene into information-rich tokens that enable context-aware 3D reasoning within the language model.

3.2. MSDec as a Unified Interface

Beyond multi-scale feature integration, MSDec serves as a unified interface that supports both human prompting, enabling interactive 3D scene understanding, and segmentation mask decoding, which extends text generation to spatially grounded 3D prediction. The following sections detail how MSDec operates in each of these modes.

User-Input Prompting. A robust 3D VLM should support user-input prompts such as clicks, boxes, and masks. ND-Tokenizer3D handles these via MSDec, aligning user intent with 3D context for interactive understanding. Concretely, inside the interactive module, we first convert user input into a binary mask $m_u \in \mathbb{R}^{N_R \times 1}$ over the finest-scale features $\mathbf{F}_R$, and the masked region is average-pooled to yield the prompt feature $\mathbf{F}_R^P$. This feature initializes an additional query $\mathbf{Q}_1^P$, concatenated with the original query set $\mathbf{Q}_1$ and jointly processed through all MSDec layers. The resulting prompt-conditioned output $\mathbf{Q}_R^P$ is then projected through the same f_{mm} to form a guidance token $\mathbf{E}_P$. By incorporating user-input prompts directly at the decoder level, MSDec effectively aligns user intentions with the multi-scale 3D context, enabling flexible, interaction-driven 3D understanding while maintaining spatial and semantic coherence.

Segmentation Mask Decoding. Beyond user prompting, MSDec also supports segmentation mask decoding as part of its unified design. When instructed to perform a segmentation task, the LLM generates a special token `[SEG]`. The

hidden state of the [SEG] token $\mathbf{H}^S$ is projected through a segmentation head f_{s} to form a query $\mathbf{Q}_1^S$, which MSDec decodes into a segmentation-aware representation $\mathbf{Q}_R^S$ using the same decoding procedure as in the previous section. A mask head f_{m} then transforms $\mathbf{Q}_R^S$ into a kernel that interacts with the finest-scale features via a dot product to predict the final 3D mask $\mathbf{M} \in \mathbb{R}^{N_R \times 1}$. This design allows segmentation to emerge naturally from the language reasoning process, rather than being treated as a separate prediction branch or a post-processing step, thereby maintaining architectural coherence and modality alignment within the unified framework.

3.3. Multimodal Token Integration

After our scene tokenization pipeline and the multi-purpose decoding interface, our system produces a set of visual tokens that serve as the visual context for the language model. The final outputs of MSDec include two types of visual tokens: (1) the scene tokens $\mathbf{E}_V$ representing holistic 3D scene information, and (2) the guidance token $\mathbf{E}_P$ derived from user-interactive cues when available. The textual instructions are tokenized using the pretrained LLM tokenizer to obtain $\mathbf{E}_T$. All tokens are concatenated to form the multimodal input sequence for the LLM, yielding

$$\hat{a} = \text{LLM}([\mathbf{E}_V; \mathbf{E}_P; \mathbf{E}_T]), \ \texttt{[SEG]} \subset \hat{a}, \tag{5}$$

where $\hat{a}$ denotes the generated textual response. The LLM can generate the special token [SEG] within the response; when present, it triggers the mask decoding process described in Sec. 3.2.

3.4. Training Strategy

Stage 1: Pre-Training 3D Encoder and MSDec. Since there are no pre-trained weights designed for NDT representations, we first pre-train the 3D Encoder and the MSDec jointly on the 3D Instance Segmentation task. To this end, we append two heads after MSDec: (1) a multi-class classification head for predicting the semantic category of each instance, and (2) a mask head f_{m}, identical to that described in Sec. 3.2, for generating instance masks. We use a combination of cross-entropy loss $\mathcal{L}_{cls}$ for classification and a segmentation loss $\mathcal{L}_m$ for mask prediction, where $\mathcal{L}_m$ consists of binary cross-entropy and Dice loss [40]. To further encourage the encoder to produce semantically aligned 3D features, we incorporate 2D visual-language supervision. Specifically, we extract 2D image features using CLIP [44] image encoder and lift them into the 3D space following the projection procedure described in Eq. (2), obtaining $\mathbf{F}_r^{\text{C}} \in \mathbb{R}^{N_r \times d_f}$. Then, we apply a cosine similarity loss between $\mathbf{F}_r^{\text{C}}$ and $\mathbf{F}_r$ as

$$\mathcal{L}_s(\mathbf{F}_r^{\text{C}}, \mathbf{F}_r) = \frac{1}{N_r} \sum_{j=1}^{N_r} 1 - \frac{\mathbf{F}_r^{j,\text{C}} \mathbf{F}_r^j}{||\mathbf{F}_r^{j,\text{C}}||_2 ||\mathbf{F}_r^j||_2}. \tag{6}$$

The final loss for Stage 1 is then given by

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_m + \lambda_2 \mathcal{L}_s(\mathbf{F}_r^{\text{C}}, \mathbf{F}_r). \tag{7}$$

Stage 2: Instruction Tuning. In this stage, we freeze 3D Encoder and MSDec while training only the projection layers, *i.e.* f_{mm} and f_{s}, along with the LLM. We conduct multi-task training using a combined dataset comprising Referring Segmentation, Visual Question Answering, and Dense Captioning tasks. Besides the next-token generation cross-entropy loss $\mathcal{L}_t$, we include the same mask prediction loss $\mathcal{L}_m$ as described in Stage 1. Furthermore, to encourage the LLM to produce semantically consistent responses, we introduce a cosine similarity loss between the hidden states $\mathbf{H}^{\hat{a}}$ of the predicted answer $\hat{a}$ and the text embeddings $\mathbf{H}^a$ of the ground-truth answer a extracted by the CLIP [44] text encoder, as defined in Eq. (6). The overall objective is then given by

$$\mathcal{L} = \mathcal{L}_t + \lambda_3 \mathcal{L}_m + \lambda_4 \mathcal{L}_s(\mathbf{H}^{\hat{a}}, \mathbf{H}^a). \tag{8}$$

4. Experiments

Datasets. We conduct experiments on the ScanNet dataset [13], including 1201 training scenes and 312 validation scenes. In Stage 1, we utilize the instance mask annotations from ScanNet200 [46], an extended version of ScanNet that introduces more fine-grained semantic categories for 3D Instance Segmentation pre-training. In Stage 2, we perform multi-task instruction tuning using a combined dataset of approximately 295k instruction–response pairs. Specifically, we train on ScanRefer [5], Nr3D [1], and Multi3DRefer [63] for Referring Segmentation, where the model localizes target objects based on textual descriptions, and evaluate on the Multi3DRefer validation set. For Visual Question Answering, which requires reasoning over 3D scenes to answer natural language questions, we train on ScanQA [2] and SQA3D [38], and evaluate on their respective validation and test sets. Lastly, for Dense Captioning, which aims to generate region- and object-level descriptions within a 3D scene, we train on Nr3D [1] and Scan2Cap [11], and evaluate on the Scan2Cap validation set. Additionally, we evaluate hallucination performance on the 3D-POPE benchmark [57], which measures a model's ability to correctly judge the presence of objects within a 3D scene.

Metrics. We use mean intersection over union (mIoU) as the evaluation metric [15] for Referring Segmentation, and report extract match accuracy (EM) and its refined version (EM-R) for SQA3D. Following prior works [15, 21, 64], we adopt CiDEr (C) [50], BLEU-4 (B-4) [41], METEOR (M) [3], and Rouge-L (R) [32] to evaluate the quality of generated text responses on ScanQA and Scan2Cap.

Implementation Details. We use Point Transformer v3 (PTv3) [54] as the 3D Encoder and employ FlashAttention-2 [14] to enhance memory and computational efficiency

Generalist 3D VLMs	Multi3DRefer	ScanQA				SQA3D		Scan2Cap			
	mIoU	C	B-4	M	R	EM	EM-R	C@0.5	B-4@0.5	M@0.5	R@0.5
Spatial 3D-LLM [52]	-	82.5	13.9	16.8	39.1	46.2	-	72.2	34.6	23.1	54.3
LEO [22]	-	101.4	13.2	20.0	49.2	50.0	52.4	72.4	**38.2**	27.9	**58.1**
Scene-LLM [16]	-	80.0	11.7	15.8	35.9	53.6	-	-	-	-	-
Chat-Scene [21]	-	87.7	14.3	18.0	41.6	**54.6**	**57.5**	77.2	36.4	**28.0**	**58.1**
Grounded 3D-LLM [10]	-	72.7	13.4	-	-	-	-	70.6	35.5	-	-
Oryx MLLM [34]	-	74.3	-	15.3	38.4	-	-	-	-	-	-
LSceneLLM [64]	-	88.2	-	18.0	40.8	-	-	-	-	-	-
3D-LLaVA [15]	42.7	92.6	**17.1**	18.4	43.1	54.5	56.6	78.8	36.9	27.1	57.7
NDTokenizer3D (Ours)	**46.0**	**98.6**	17.0	**19.4**	**44.9**	54.4	57.1	**79.0**	36.7	27.1	57.7

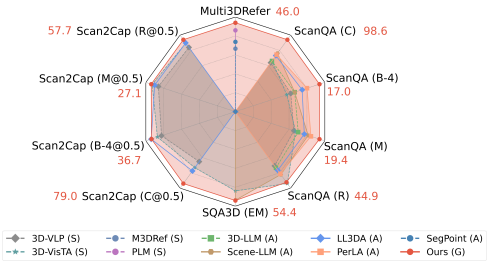

Table 1. **Left:** Quantitative comparison of state-of-the-art generalist 3D VLMs. Results for LEO [22] on ScanQA are shown in gray as it operates under a different setting with access to ground-truth objects and are thus not directly comparable. The best and second-best results for each metric are **bold** and underlined, respectively. NDTokenizer achieves the strongest overall performance, particularly on segmentation and QA tasks. **Right:** Radar plot comparing the per-task performance of specialist models and adapted 3D VLMs. NDTokenizer consistently surpasses both categories across all tasks.

within MSDec. MSDec takes 850 initial learnable queries. The multimodal alignment head, the segmentation head, and the multi-class classification head are implemented as two-layer MLPs. For the LLM endpoint, we adopt the language model from the LLaVA-1.5-7B model [33] in our framework. Training is conducted on 4× NVIDIA A100 GPUs with the DeepSpeed toolkit [45] for distributed optimization. We apply LoRA fine-tuning [19] on the LLM. We perform instruction tuning for one epoch. The batch size is set to 32 per GPU. We use the AdamW optimizer [36] with a cosine annealing learning rate schedule, starting from an initial learning rate of 8×10^{-5} and a warm-up ratio of 10%. Additional implementation details are provided in the Supplementary Materials.

4.1. Quantitative Comparisons

We compare NDTokenizer3D with existing approaches in Tab. 1. The models compared are categorized into three groups: specialist models (S), adapted 3D VLMs (A), and generalist 3D VLMs (G). Specialist models are designed for specific tasks and lack generalization, while adapted 3D VLMs require task-specific fine-tuning that improves in-domain performance but limits cross-task transfer. In contrast, generalist 3D VLMs are trained on unified multi-task datasets, enabling a single model to perform diverse 3D vision–language tasks. The left panel of Tab. 1 summarizes the comparison among generalist 3D VLMs, while the right shows results for specialist and adapted models. Comparison with more methods and full metrics is provided in the Supplementary Materials.

3D Referring Segmentation. Multi3DRefer [63] benchmark measures a model's ability to localize a flexible number of objects in real-world 3D scenes from natural language descriptions by predicting 3D masks. Each description in Multi3DRefer may refer to zero, one, or multiple target objects. As shown in Tab. 1, only our NDTokenizer3D and 3D-LLaVA [15] are capable of performing both language-centric tasks and point-level understanding tasks (*e.g.* Referring Segmentation). Our NDTokenizer3D achieves 46.0 mIoU, surpassing 3D-LLaVA by **+3.3**. This improvement stems from two key design choices: (1) our multi-scale NDT representation preserves both local geometric details and global contextual information, whereas 3D-LLaVA relies on superpoint-pooling and token selection, effectively a form of downsampling; and (2) our MSDec integrates features across multiple scales, unlike 3D-LLaVA's single-scale design. Together, these components enable more spatially coherent scene representations and more precise language-grounded segmentation.

3D Visual Question Answering. ScanQA [2] evaluates a model's 3D perception and reasoning capabilities across diverse question types, including object counting, identification, attribute recognition, and appearance reasoning. In contrast, SQA3D [38] focuses on situated reasoning, requiring the model to infer spatial context, such as position, orientation, and viewpoint from text and reason about surrounding 3D environments. Our NDTokenizer3D consistently achieves top performance on both datasets. Notably, on ScanQA, it attains 98.6 CiDEr and 19.4 METEOR, surpassing the second-best results by **+6.0** and **+1.0**, respectively. These results validate the strong reasoning and alignment capabilities of our proposed 3D scene tokenization pipeline when integrated with LLMs.

3D Dense Captioning. The Scan2Cap benchmark [11] evaluates a model's ability to describe a user-specified object's appearance and its spatial relations with surrounding objects. Following prior works [15, 22, 65], we adopt object proposals generated by Mask3D [48] for evaluation, while notably not using Mask3D outputs during training. Our NDTokenizer3D treats these object proposals as visual prompts, which are processed by MSDec (Sec. 3.2) to provide guided visual context to the LLM. Our NDTokenizer3D achieves competitive results across all metrics, demonstrating strong capability in generating spatially coherent and contextually rich descriptions grounded in 3D environments. Moreover, this experiment highlights our model's capability to interpret and incorporate high-level human intent directly into its reasoning process. This prop-

erty is essential for achieving interactive 3D understanding, indicating that NDTokenizer3D goes beyond passive scene perception to support intent-driven multimodal reasoning.

Methods	Random		Popular		Adversarial	
	Prec	Acc	Prec	Acc	Prec	Acc
LEO [22]	51.95	52.91	48.30	47.27	48.47	47.52
3D-LLaVA [15]	73.24	80.32	66.86	74.11	63.13	69.88
Ours	**80.34**	**84.12**	**69.69**	**75.51**	**66.15**	**72.03**

Table 2. Hallucination performance on 3D-POPE. Our NDTokenizer3D achieves the lowest hallucination rates across all settings.

Hallucination Performance. 3D-POPE [57] consists of yes/no questions about whether a specific object exists in the scene. This benchmark is balanced, containing an equal number of positive and negative questions. To construct negative questions (nonexistent objects), 3D-POPE provides three different settings: (1) Random Sampling: randomly selects an object category that is absent in the current scene. This setting evaluates whether the model tends to hallucinate arbitrary objects without contextual grounding. (2) Popular Sampling: selects the most frequent object categories across the dataset that are not present in the current scene. This tests whether the model hallucinates common objects that it has seen often during training. (3) Adversarial Sampling: for each object that truly exists in the scene, it identifies objects that often co-occur with it in the training data but are absent in the current scene. The most strongly co-occurring absent object is then used as an adversarial negative sample. This design targets realistic contextual hallucinations rather than random or popular object confusions. As shown in Tab. 2, our model exhibits substantially lower hallucination rates across all three settings, consistently outperforming previous generalist methods by a large margin. These results show that NDTokenizer3D provides more informative scene tokens that are faithful to the underlying geometry, leading to more accurately grounded LLM predictions.

4.2. Ablation Studies

Comparison with Downsampling. We conduct an ablation study comparing our proposed multi-scale NDT compression with a naive multi-scale downsampling baseline. Specifically, we replace the NDT cells with voxel-downsampled point clouds containing the same number of points as NDT cells at each scale, while keeping all other architectural components identical. Formally,

$$\mathbf{P}_r = \left\{P_r^j\right\}_{j=1}^{N_r} \in \mathbb{R}^{N_r \times 6},\ P_r^j = \left[p_r^j; c_r^j\right], \tag{9}$$

where p_r^j and c_r^j denote the 3D coordinates and RGB color, respectively. The resulting multi-scale point sets $\{\mathbf{P}_r\}_{r=1}^{R}$ are processed by the same tokenization pipeline and fed into the LLM for reasoning. Results are summarized in Tab. 3. Our method consistently outperforms the downsampled baseline across all tasks, confirming that NDT provides richer and more discriminative scene representations. On ScanQA in particular, replacing NDT with downsampled points leads to a noticeable drop, indicating that naive downsampling removes fine-grained geometric cues essential for correct reasoning. Interestingly, the baseline still performs competitively on Referring Segmentation, surpassing 3D-LLaVA [15] (Tab. 1), highlighting that multi-scale integration is highly beneficial for point-level understanding. A similar trend appears on ScanQA, where the baseline also outperforms 3D-LLaVA, showing that the multi-scale integration helps capture inter-object relations and broader contextual cues, even without NDT statistics.

Methods	Multi3DRefer	ScanQA				SQA3D		Scan2Cap
	mIoU	C	B-4	M	R	EM	EM-R	C@0.5
Baseline	45.3	94.7	16.2	18.8	43.9	53.9	56.5	78.1
Ours	**46.0**	**98.6**	**17.0**	**19.4**	**44.9**	**54.4**	**57.1**	**79.0**

Table 3. Ablation comparing multi-scale NDT compression with a downsampling baseline. Best socres are in **bold**. Our NDT-based tokenization consistently improves performance across all tasks.

Scales	Multi3DRefer	ScanQA		SQA3D		Scan2Cap	
	mIoU	C	B-4	EM	EM-R	C@0.5	B-4@0.5
$r = 3$	40.1	91.8	15.1	51.1	53.7	77.0	35.9
$r = 4$	44.9	96.6	16.6	54.3	56.9	76.6	35.7
$r = \{2, 3\}$	41.3	90.6	15.2	50.5	53.3	75.0	35.9
$r = \{3, 4\}$	**46.2**	94.4	15.9	53.9	56.5	77.4	35.8
$r = \{2, 3, 4\}$	46.0	**98.6**	**17.0**	54.4	57.1	**79.0**	36.7
$r = \{1, 2, 3, 4\}$	44.2	94.7	16.3	**54.6**	**57.3**	77.1	35.9

Table 4. Ablation comparing number of scales. Best scores are in **bold**. Three-scale variant provides a balance between detail preservation and stable reasoning.

Number of Scales. We study the impact of the number of scales R used in our multi-scale representation. As shown in Tab. 4, we compare single-scale, two-scale, three-scale, and four-scale variants. The three-scale configuration yields the strongest overall performance. Single- and two-scale models lack sufficient cross-scale context, limiting their ability to capture both fine geometry and global structure. Increasing to four scales introduces overly fine partitions that contribute noise and lead to mild overfitting, while also incurring higher computational cost. These results confirm that three scales provide an effective balance between detail preservation and stable reasoning. More comparisons are provided in Supplementary Materials.

Number of Queries. Tab. 5 reports the impact of varying the number of learnable queries. Increasing the query count generally improves performance, as a larger query set provides greater expressive capacity and allows MSDec to aggregate richer multi-scale information. Performance saturates around 400–850 queries, indicating that MSDec captures sufficient scene information within this range. Beyond

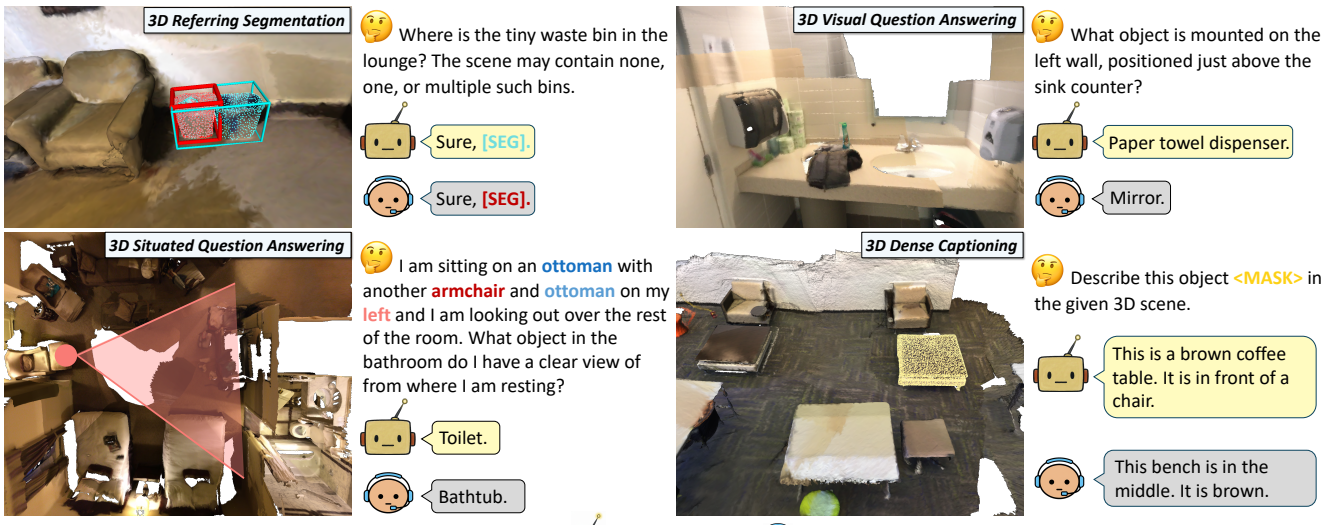

Figure 3. Qualitative comparison between NDTokenizer3D and 3D-LLaVA across four tasks, showing NDTokenizer3D's improved grounding, spatial reasoning, and object understanding.

Number of Queries	Multi3DRefer mIoU	ScanQA C	ScanQA B-4	SQA3D EM	SQA3D EM-R	Scan2Cap C@0.5	Scan2Cap B-4@0.5
100	44.3	90.2	14.6	53.1	55.8	77.1	35.9
200	45.2	94.6	16.4	54.2	56.8	**79.2**	**36.7**
400	**46.0**	94.8	16.2	54.0	56.6	77.4	36.3
850	**46.0**	**98.6**	**17.0**	54.4	57.1	79.0	**36.7**
1200	**46.0**	96.3	16.0	54.8	57.3	78.5	36.2
1600	45.9	95.9	16.5	**54.9**	**57.5**	77.3	36.4

Table 5. Ablation comparing number of queries.

this point, adding more queries yields diminishing returns and can even slightly degrade results, likely due to overfitting to nuanced noise. We adopt 850 queries as it offers the best balance across all benchmarks.

4.3. Result Visualizations

We present comparison between our NDTokenizer and 3D-LLaVA [15] in Fig. 3 across four tasks.

Segmentation. 3D-LLaVA often fails to localize all target instances, indicating that its superpoint-pooling–based (downsampling-like) representation loses the fine-grained cues necessary to distinguish nearby or similarly shaped objects. In contrast, NDTokenizer3D accurately segments all referenced instances, benefiting from the preserved local details in our multi-scale NDT representation.

VQA. 3D-LLaVA tends to rely heavily on language priors rather than the actual 3D scene. In this example, it hallucinates the answer "mirror", a plausible but incorrect guess given typical sink layouts. NDTokenizer3D instead predicts correctly by grounding its reasoning in the visual tokens.

SQA. From the described viewpoint (marked by the red dot), the bathtub is occluded by the wall, leaving the toilet as the only visible candidate. 3D-LLaVA misidentifies the object as its downsampled representation lacks sufficient spatial context. NDTokenizer3D, leveraging information-rich multi-scale features, correctly reasons about visibility and object–environment relationships.

Captioning. 3D-LLaVA misrecognizes the object and its spatial relation to the environment, again reflecting the limitations of its scene encoding. NDTokenizer3D instead produces correct and spatially coherent descriptions, grounded in the detailed multi-scale representation.

Overall, these qualitative results show that NDTokenizer3D provides more reliable scene representations, capturing fine geometry and broader structure, and achieves more accurate, consistent results across diverse 3D tasks.

5. Conclusion

We present NDTokenizer3D, a general-purpose 3D VLM capable of handling a broad range of 3D scene understanding tasks. Our approach introduces a novel three-stage scene tokenization pipeline: (1) constructing a multi-scale Normal Distributions Transform (NDT) representation from high-resolution point clouds, (2) extracting multi-scale 3D features, and (3) aggregating them through the Multi-Scale NDT Decoder (MSDec) to produce holistic scene tokens. The multi-scale NDT preserves both fine-grained geometric details and global contextual information from raw point clouds without downsampling, forming a compact yet expressive representation of 3D scenes. Beyond feature aggregation, MSDec serves as a unified interface that incorporates human prompts and supports segmentation mask decoding, thereby bridging language-level reasoning with spatial understanding within a single architecture. Across four 3D vision–language benchmarks, NDTokenizer3D demonstrates impressive performance and broad task coverage, highlighting its effectiveness and versatility in advancing general-purpose 3D vision–language understanding.

Acknowledgement

This work was conducted during Yutao's internship at Microsoft, which also provided computational and research support.

References

[1] 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*, pages 422–440. Springer, 2020. 2, 5

[2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 2, 5, 6

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5

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

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 2, 5

[6] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. 2021. 3

[7] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11124–11133, 2023. 2

[8] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26428–26438, 2024. 2

[9] Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, and Tao Chen. Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7331–7347, 2024. 2, 3

[10] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. 2, 3, 6

[11] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 2, 5, 6

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

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5

[14] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 5

[15] 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*, pages 3772–3782, 2025. 2, 5, 6, 7, 8

[16] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2195–2206. IEEE, 2025. 2, 3, 6

[17] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pages 349–367. Springer, 2024. 2

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

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

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

[21] 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, 2024. 2, 5, 6

[22] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 3, 6, 7

[23] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d

segmentation via large language model. In *International Conference on 3D Vision 2025*, 2025. 2

[24] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 2

[25] Zhuoxu Huang, Mingqi Gao, and Jungong Han. Point linguist model: Segment any object via bridged large 3d-language model. *arXiv preprint arXiv:2509.07825*, 2025. 2

[26] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. In *European Conference on Computer Vision*, pages 528–545. Springer, 2022. 3

[27] Bu Jin, Yupeng Zheng, Pengfei Li, Weize Li, Yuhang Zheng, Sujie Hu, Xinyu Liu, Jinwei Zhu, Zhijie Yan, Haiyang Sun, et al. Tod3cap: Towards 3d dense captioning in outdoor scenes. In *European Conference on Computer Vision*, pages 367–384. Springer, 2024. 2

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

[29] Minjung Kim, Hyung Suk Lim, Soonyoung Lee, Bumsoo Kim, and Gunhee Kim. Bi-directional contextual attention for 3d dense captioning. In *European Conference on Computer Vision*, pages 385–401. Springer, 2024. 2

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[31] Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, Xiangde Liu, and Rong Wei. 3dmit: 3d multi-modal instruction tuning for scene understanding. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–5. IEEE, 2024. 3

[32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 6

[34] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 6

[35] Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. Reasongrounder: Lvlm-guided hierarchical feature splatting for open-vocabulary 3d visual grounding and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3718–3727, 2025. 3

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[37] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. *Advances in neural information processing systems*, 37:68803–68832, 2024. 3

[38] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 2, 5, 6

[39] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2024. 2

[40] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5

[41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[42] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5607–5612, 2023. 2

[43] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4551–4559, 2024. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[45] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020. 6

[46] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European conference on computer vision*, pages 125–141. Springer, 2022. 5

[47] Pranav Saxena, Avigyan Bhattacharya, Ji Zhang, and Wenshan Wang. Llm-rg: Referential grounding in outdoor scenarios using large language models. *arXiv preprint arXiv:2509.25528*, 2025. 2

[48] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 6

[49] Eijiro Takeuchi and Takashi Tsubouchi. A 3-d scan matching using improved 3-d normal distributions transform for mobile robotic mapping. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3068–3073. IEEE, 2006. 2, 3

[50] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[51] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022. 3

[52] Xiaoyan Wang, Zeju Li, Yifan Xu, Jiaxing Qi, Zhifei Yang, Ruifei Ma, Xiangde Liu, and Chao Zhang. Spatial 3d-llm: Exploring spatial awareness in 3d vision-language models. *arXiv preprint arXiv:2507.16524*, 2025. 6

[53] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5940–5948, 2024. 2

[54] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024. 5

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

[56] Dejie Yang, Zhu Xu, Wentao Mo, Qingchao Chen, Siyuan Huang, and Yang Liu. 3d vision and language pre-training with large-scale synthetic data. *arXiv preprint arXiv:2407.06084*, 2024. 3

[57] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29501–29512, 2025. 5, 7

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

[59] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021. 2

[60] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022. 2

[61] Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. Agent3d-zero: An agent for zero-shot 3d understanding. In *European Conference on Computer Vision*, pages 186–202. Springer, 2024. 3

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

[63] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023. 2, 5, 6

[64] Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3761–3771, 2025. 2, 5, 6

[65] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 3, 6