# 3D-DRES: Detailed 3D Referring Expression Segmentation

**Qi Chen**[*1], **Changli Wu**[*1 2], **Jiayi Ji**[†1 3], **Yiwei Ma**[1], **Liujuan Cao**[1]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]Shanghai Innovation Institute
[3]National University of Singapore
chenqi@stu.xmu.edu.cn, wuchangli@stu.xmu.edu.cn, jjyxmu@gmail.com, mayiwei@stu.xmu.edu.cn, caoliujuan@xmu.edu.cn

## Abstract

Current 3D visual grounding tasks only process sentence-level detection or segmentation, which critically fails to leverage the rich compositional contextual reasonings within natural language expressions. To address this challenge, we introduce Detailed 3D Referring Expression Segmentation (3D-DRES), a new task that provides a phrase to 3D instance mapping, aiming at enhancing fine-grained 3D vision-language understanding. To support 3D-DRES, we present DetailRefer, a new dataset comprising 54,432 descriptions spanning 11,054 distinct objects. Unlike previous datasets, DetailRefer implements a pioneering phrase-instance annotation paradigm where each referenced noun phrase is explicitly mapped to its corresponding 3D elements. Additionally, we introduce DetailBase, a purposefully streamlined yet effective baseline architecture that supports dual-mode segmentation at both sentence and phrase levels. Our experimental results demonstrate that models trained on DetailRefer not only excel at phrase-level segmentation but also show surprising improvements on traditional 3D-RES benchmarks.

**Github** — https://github.com/80chen86/3D-DRES

## 1 Introduction

Vision-language integration stands at the forefront of computer vision research, enabling machines to understand and reason about the visual world through natural language (Antol et al. 2015; Johnson et al. 2017; Krishna et al. 2017; Chen et al. 2015; Dang et al. 2025a,b). In recent years, with the rapid advancement of 3D sensing technologies and deep learning models, 3D vision-language tasks have emerged as a prominent research focus in the community (Yuan et al. 2022; He et al. 2025b,a; Azuma et al. 2022; Wu et al. 2024c; Chen, Chang, and Nießner 2020; Huang et al. 2021). These tasks enable critical applications across robotics, autonomous navigation, mixed reality, and assistive technologies, where understanding the relationship between 3D environments and natural language is essential.

Among these, 3D visual grounding tasks are particularly crucial as they require localizing targets in 3D scenes based on textual instructions—a fundamental capability for embodied AI and autonomous systems. The field has evolved systematically, beginning with 3D-REC (Chen, Chang, and Nießner 2020; Guo et al. 2025; Mi et al. 2025; Wang et al. 2023b; Shi, Wu, and Lee 2024; Zhu et al. 2024; Li et al. 2024b; Peng, Zheng, and Huang 2025), which localizes objects using coarse 3D bounding boxes and formulates the problem as coordinate regression (Fig. 1-(a)). Subsequently, 3D-RES was introduced to address the need for finer-grained localization (Huang et al. 2021; He and Ding 2024; Qian et al. 2024; Liu et al. 2024), requiring point-level segmentation and transforming the task into an expression-to-point matching problem (Fig. 1-(b)). However, both tasks were constrained by their ability to handle only one-to-one mappings between sentences and objects, limiting their practical applicability in scenarios where instructions might refer to multiple objects or none at all. To address this limitation, 3D-GRES (Fig. 1-(c)) expanded the task formulation to accommodate zero, one, or multiple targets per textual description (Wu et al. 2024a; Chen et al. 2025; Zhang, Gong, and Chang 2023; Wang et al. 2025b). These developments have significantly advanced the field of 3D visual grounding, establishing it as a cornerstone of 3D scene understanding.

Despite these significant advancements, existing 3D visual grounding tasks still suffer from a prominent issue: the single-unit assumption (we define an "unit" as one or more objects that do not need to be distinguished). Specifically, current tasks are limited to sentence-level segmentation or localization, meaning they focus only on a single unit described in a sentence, which greatly limits both practical applications and academic research. In real-world applications, completing user-issued commands often requires attention to all units mentioned. For example, in the common instruction, "Put these clothes into the washing machine", both "clothes" and "washing machine" are essential units to consider. In academic research, the single-unit assumption hinders the comprehensive evaluation of models' fine-grained linguistic understanding and constrains the modeling of intra-sentence relationships and semantic structures. As illustrated in Fig. 1-(d), traditional 3D-RES approaches provide no mechanism to determine whether a model correctly comprehends individual elements like "table" and "TV", even when localization appears successful

[*]These authors contributed equally.
[†]Corresponding author

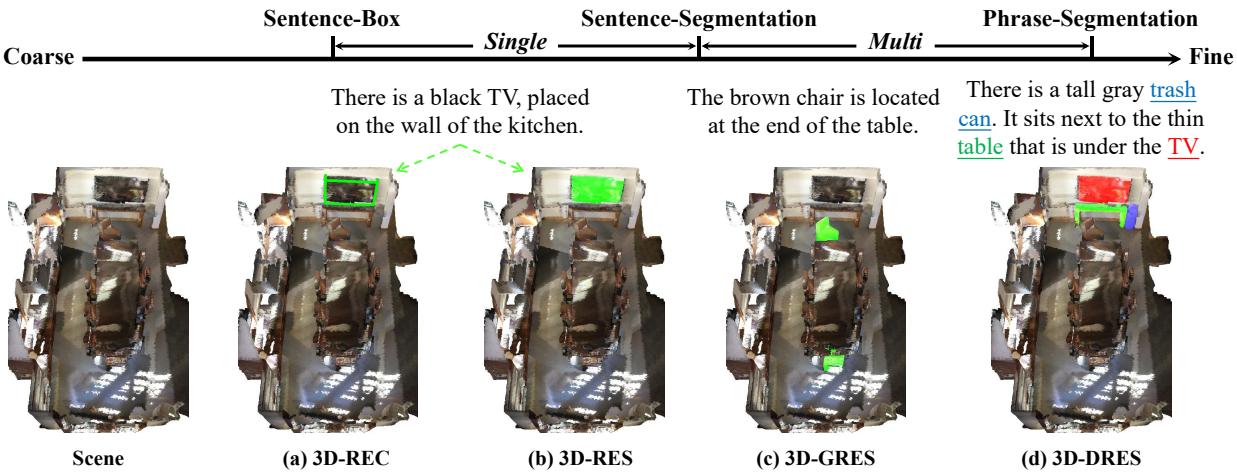


Figure 1: Illustration of 3D visual grounding tasks. (a) 3D Referring Expression Comprehension (3D-REC). (b) 3D Referring Expression Segmentation (3D-RES). (c) Generalized 3D Referring Expression Segmentation (3D-GRES). (d) Detailed 3D Referring Expression Segmentation (3D-DRES).

for "trash can". This limitation creates a critical gap in interpretability, as effective referring expression understanding inherently depends on contextual reasoning capabilities within the text. The need for a more granular, phrase-level understanding that would enable 3D vision-language models to develop robust contextual reasoning capabilities represents a significant unexplored opportunity in this domain.

Accordingly, we propose a novel task, Detailed 3D Referring Expression Segmentation (3D-DRES). This task requires models to focu all units within sentences. Specifically, as shown in Fig.1-(d), the dataset provides the positions of all units to be segmented in the sentence (such as "trash can", "table", and "TV"), and the model needs to generate the corresponding mask for each unit separately. The nature of the 3D-DRES task format inherently determines that models trained under this task focus more on the fine-grained semantic within sentences and is more aligned with practical applications. Meanwhile, the 3D-RES and the 3D-DRES can be mutually beneficial. We demonstrate this point with experiments in the subsequent sections.

A critical challenge in advancing this new paradigm is the absence of a suitable dataset. Creating such a resource is particularly challenging given the substantial costs associated with 3D annotation. To break through this dilemma, we invested approximately 600 hours to build DetailRefer, a new dataset built upon ScanRefer (Chen, Chang, and Nießner 2020) through a combination of meticulous manual annotation and large language model (Yang et al. 2024a) assistance. DetailRefer contains 54,432 descriptions covering 11,054 distinct objects, with an average text length of 24.9 tokens - significantly exceeding existing datasets (9.7-20.1 tokens). Unlike the sentence-segmentation annotation format prevalent in current datasets (Chen, Chang, and Nießner 2020; Achlioptas et al. 2020; Zhang, Gong, and Chang 2023) where one sentence corresponds to one segmentation result, DetailRefer implements a pioneering phrase-segmentation format where each noun phrase corresponds to a distinct segmentation mask. This results in an unprecedented density of 2.9 masks per text. The dataset strategically incorporates 7.4% "Long" texts (exceeding 50 tokens) and numerous "Complex" samples requiring segmentation of four or more noun phrases, establishing a robust framework for evaluating models' fine-grained linguistic understanding capabilities in 3D environments.

We observe that most existing models (Huang et al. 2021; Chen et al. 2025; Qian et al. 2024; Xu et al. 2025) are designed for "sentence-segmentation" sample formats and are unable to output multiple masks or specify masks for particular tokens. This limitation makes it impossible to apply current methods to the 3D-DRES task directly. Therefore, as initiators of this task, we propose DetailBase, a purposefully streamlined but effective baseline to establish the foundation for future research. Our design philosophy prioritizes simplicity to ensure high scalability and adaptability, while simultaneously demonstrating sufficient effectiveness to validate the task's potential. DetailBase features an elegant architecture that supports both sentence-level and phrase-level segmentation, with our comprehensive experiments confirming its efficacy across various evaluation dimensions. Importantly, our results demonstrate that training on this fine-grained task yields surprising improvements even on traditional 3D-RES benchmarks, suggesting that phrase-level understanding enhances overall spatial reasoning capabilities.

To sum up, our main contributions are as follows:

- We introduce 3D-DRES, a novel fine-grained visual grounding task that is designed to enhance the ability to understand and localize textual context in 3D vision and language tasks.
- We have created a new dataset based on the Scannet indoor point cloud scenes, combining human annotations with large language models.
- We provide a simple, yet effective framework to serve as a foundational starting point for studying the 3D-DRES.

## 2 Related Work

### 2.1 3D Referring Expression Segmentation

The 3D-RES task was first proposed by Huang et al. (Huang et al. 2021) in 2021, aiming to segment the target object specified in the point cloud based on a given textual description. There are currently three main datasets for this task, namely ScanRefer (Chen, Chang, and Nießner 2020), Sr3D, and Nr3D (Achlioptas et al. 2020). The text descriptions in these three datasets are all based on the indoor point cloud scene dataset Scannet (Dai et al. 2017). In terms of details, the descriptions in ScanRefer and Nr3D are manually annotated by humans. Sr3D is a template based description generated based on the mutual positional relationship between objects, which has lower complexity.

In response to the shortcomings of 3D-RES tasks, namely the existence and uniqueness of target objects, Wu et al. (Wu et al. 2024a) proposed the 3D-GRES task, aimed at lifting the limitation on the number of target objects. The task currently only has one dataset, Multi3DRefer (Zhang, Gong, and Chang 2023), proposed by Zhang et al. It modifies and enhances the text based on ScanRefer, so that the number of targets corresponding to the text is no longer fixed.

To address these two tasks, researchers have proposed numerous methods (He and Ding 2024; He et al. 2024; Xu et al. 2025; Huang et al. 2021; Qian et al. 2024; Wu et al. 2024a) over time. However, most of these approaches are unfortunately limited to sentence-level segmentation, with their frameworks barely supporting fine-grained segmentation of noun phrases within sentences. Even when we extend our view to approaches (Chen et al. 2022; Feng et al. 2021; Yuan et al. 2021; Yang et al. 2023b; Luo et al. 2022; Wang et al. 2025a; Wu et al. 2023; Yang et al. 2021; Jain et al. 2022; Zhao et al. 2021) designed for the 3D-REC task, the situation does not improve much. Therefore, in this paper, we design a simple yet effective framework that supports both sentence-level and phrase-level segmentation.

### 2.2 2D Referring Expression Segmentation

Compared to 3D-RES, 2D-RES (Yang et al. 2022; Lai et al. 2024; Chng et al. 2024; Yang et al. 2024b; Wang et al. 2025c) was proposed and studied by Hu et al. (Hu, Rohrbach, and Darrell 2016) as early as 2016, with the goal of segmenting target objects in images based on given text descriptions. The RES dataset in the 2D domain is more abundant compared to the 3D domain, with classic examples including RefCOCO, RefCOCO+ (Kazemzadeh et al. 2014), and RefCOCOg (Mao et al. 2016). Similarly, there is also GRES in the 2D field, with its dataset being gRefCOCO (Liu, Ding, and Jiang 2023). In addition, Flickr30k Entities (Plummer et al. 2015; Zhan et al. 2024; Wan et al. 2024; Shen et al. 2024; Zhang et al. 2024; Pham et al. 2024) and Panoptic Narrative Grounding (González et al. 2021; Li et al. 2024a; Hui et al. 2023; Wang et al. 2023a; Ding et al. 2022; Yang et al. 2023a) are two datasets that are similar to this paper, both of which require segmentation of noun phrases in sentences. But the difference is that the former's supervisory signal is a box, while the latter is a text that corresponds to a panoramic description of an image.


Figure 2: Example of DetailRefer.


Figure 3: Category distribution of phrases in DetailRefer.

## 3 DetailRefer Dataset

In this section, we first introduce the construction process of the DetailRefer (Sec. 3.1). Afterwards, introduce the relevant statistical data of the dataset and compare it with other datasets (Sec. 3.2). Finally, we introduce the 3D-DRES task definition and metrics based on this dataset (Sec. 3.3).

### 3.1 Dataset Creation

Similar to Multi3DRefer (Zhang, Gong, and Chang 2023), we modify and enhance the ScanRefer (Chen, Chang, and Nießner 2020) dataset to derive DetailRefer. In the first phase, we first use a program to divide the descriptions in ScanRefer, grouping all descriptions of the same object together. Then, we ask the LLM to consolidate these sentences into a more comprehensive new description. It is worth noting that we do not directly use the texts from ScanRefer because each text typically involves fewer objects, which aligns less with our task objectives. After obtaining approximately 10,000 new descriptions, we manually annotate all noun phrases in each text to link them to specific objects in the 3D scene. During the annotation process, any inaccuracies in the descriptions are also corrected.

After the first phase, we initially obtained a dataset that met our criteria, but it was relatively small in scale. Therefore, in the second phase, we needed to enhance and expand this dataset. Specifically, we transform the annotated text into a format that is easier for LLM to understand. This means appending a '()' immediately after each noun phrase, with the object IDs corresponding to that noun phrase enclosed within the '()'. We then fed these sentences into the

| Dataset | Avg. length | Long | Avg. mask | Num |
|---|---|---|---|---|
| ScanRefer | 20.1 | 0.5% | 1.0 | 51583 |
| Sr3D | 9.7 | 0.0% | 1.0 | 83572 |
| Nr3D | 11.5 | 0.3% | 1.0 | 41503 |
| Multi3DRefer | 15.1 | 0.0% | 1.0 | 61926 |
| DetailRefer (ours) | 24.9 | 7.4% | 2.9 | 54432 |

Table 1: Datasets comparison.

LLM, instructing it to generate several different expressions while retaining the original semantic meaning and keeping the object IDs following the corresponding noun phrases. Using this method, we expanded the dataset to five times its original size, achieving a scale comparable to ScanRefer.

At this point, we have actually completed the initial construction of the dataset. However, considering that each text currently only describes a small area centered around a single object, we aim to obtain more challenging texts that cover larger areas. Thus, we traverse each object mentioned in the dataset, extract all texts from the first-phase dataset that involve that object, and input these texts into the LLM for integration to generate texts with broader descriptions. With this, the construction of the dataset is complete, and the full construction pipeline can be found in the appendix. We have included a sample from the dataset in the Fig. 2.

### 3.2 Dataset Statistics

DetailRefer contains a total of 54,432 descriptions, involving 11,054 different objects from the Scannet (Dai et al. 2017). After dividing according to scenes in Scannet, the training, validation, and test sets contain 43,282, 5,398, and 5,752 descriptions, respectively. DetailRefer comprises 156636 noun phrases to be segmented. We present the category distribution of the objects corresponding to these phrases in Fig. 3. Additionally, among all the noun phrases to be segmented, 15,001 phrases involve multiple objects; in all texts, long texts (more than 50 tokens) account for 7.4%.

To facilitate comparison, we have compiled the various statistics of mainstream datasets (Chen, Chang, and Nießner 2020; Achlioptas et al. 2020; Zhang, Gong, and Chang 2023) in Tab 1. It can be seen that compared to other datasets, our dataset excels in average text length. Additionally, it is observed that long texts exceeding 50 tokens are extremely rare, almost non-existent, in other datasets. However, in our dataset, 7.4% of the texts are long, which better tests the model's understanding of natural language. Furthermore, since other datasets correspond to sentence-level segmentation tasks, each text only has one mask. In contrast, our dataset corresponds to noun phrase-level segmentation tasks, thus each text averages 2.9 masks. From the comparison of text quantities, it is evident that our dataset is comparable in scale to existing mainstream datasets.

### 3.3 3D-DRES Task

**Task Definition** The Detailed 3D Referring Expression Segmentation (3D-DRES) task aims to segment out masks corresponding to each target phrase given in a sentence from a point cloud scene. Specifically, first, given a point cloud scene $P \in \mathbb{R}^{N_p \times f}$, where $N_p$ represents the number of points and $f$ is the feature length (including coordinates XYZ, color RGB, etc.). Secondly, given a textual description $T \in \mathbb{R}^L$, where $L$ is the number of tokens in the text. Finally, given a set of indices $I = \{i_1, i_2, \ldots, i_k\}$, corresponding to the positions of $k$ nouns in the text that need segmentation, the model is required to output point cloud scene masks $Mask \in \mathbb{R}^{k \times N_p}$ for all nouns.

**Metrics** First, we focus on phrase-level mean IoU (mIoU), Acc@0.25, and Acc@0.5. At the phrase level, mIoU is the average of the IoUs calculated for all terms to be segmented, while Acc@0.25 and Acc@0.5 represent the proportion of segmented terms with an IoU greater than 0.25 and 0.5 out of all segmented terms. Their formulas are as follows:

$$\text{mIoU} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{K_i} IoU_{i,j}}{\sum_{i=1}^{M} K_i}, \tag{1}$$

$$\text{Acc@}t = \frac{\sum_{i=1}^{M} \sum_{j=1}^{K_i} B_t(IoU_{i,j})}{\sum_{i=1}^{M} K_i}, t \in \{0.25, 0.5\} \tag{2}$$

where $M$ represents the number of descriptions in the dataset, $K_i$ denotes the number of nouns that need to be segmented in the $i$-th description, $IoU_{i,j}$ represents the IoU value corresponding to the $j$-th noun in the $i$-th description, and $B_t(\cdot)$ indicates that it returns 1 if the parameter is greater than $t$, otherwise it returns 0. Additionally, we also focus on the sentence-level mean IoU (mIoU-S), whose calculation formula is as follows:

$$\text{mIoU-S} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{K_i} \sum_{j=1}^{K_i} IoU_{i,j}, \tag{3}$$

where the meanings of all symbols are the same as in Eq. 1. The mIoU-S reflect the model's understanding capability at the sentence level.

Finally, we also pay special attention to the model's performance on long texts and complex scene descriptions. Specifically, we define texts with more than 50 tokens as long texts, and evaluate the metrics of the model on such texts to reflect its understanding capability of long texts. Meanwhile, we define texts that require segmentation of four or more phrases as complex scene descriptions, and test the model's metrics on these texts to reflect its understanding capability of complex scene descriptions.

## 4 Baseline of 3D-DRES

Existing 3D-RES methods are based on the assumption of "sentence-level segmentation," which typically cannot output multiple masks or specify masks for particular tokens. This limitation means that they cannot be directly applied to our new task. Therefore, in this section, as the proposers of 3D-DRES, we introduce a simple, effective, and highly extensible framework to serve as a foundational starting point for investigating this task. We name this framework DetailBase, and an overview is shown in Fig. 4.

According to the task definition in Sec. 3.3, the inputs for this task include a point cloud scene $P \in \mathbb{R}^{N_p \times f}$, a textual description $T$, and the position $I$ of the noun to be

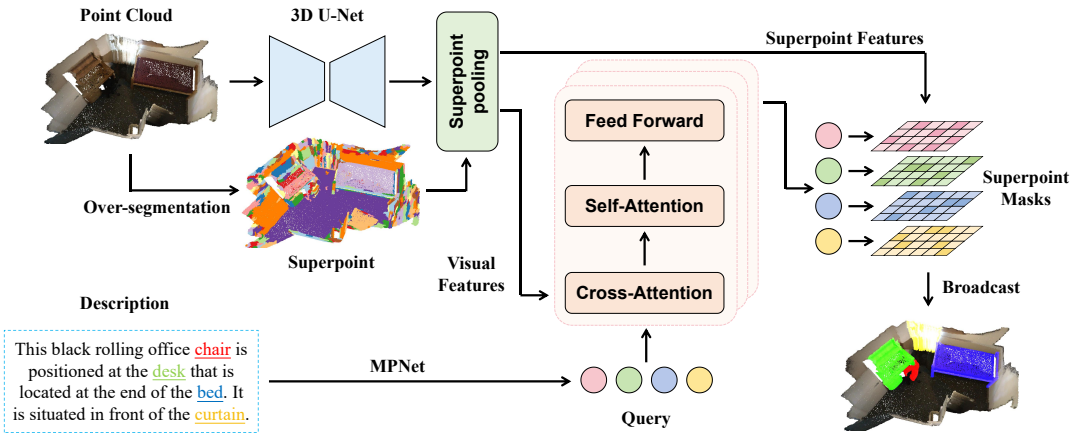


Figure 4: The overview of the Detailed 3D Referring Expression Segmentation Baseline (DetailBase).

segmented. We first obtain point-level features $F_p$ by feeding the point cloud (here, we only use coordinates XYZ and color RGB as the initial features for each point) into a 3D U-Net network (Graham, Engelcke, and Van Der Maaten 2018). Given that the number of these features is excessively large, we adopt the same approach as previous models (Wu et al. 2024b; Sun et al. 2023), namely superpoint pooling, to simplify them. Specifically, we perform an unsupervised oversegmentation on the $P$, resulting in $N_s$ superpoints $\{SP_i\}_{i=1}^{N_s}$, where $N_p >> N_s$ (Landrieu and Simonovsky 2018). Afterward, we average the features of all points belonging to the same superpoint. Finally, through two independently weighted linear transformations, the pooling features are converted into visual features for multimodal information fusion and superpoint features for predicting the mask. The entire process is formulated as follows:

$$F_{pool} = \text{SPPool}(F_p, P), \tag{4}$$

$$F_v = F_{pool}W_1, F_{sp} = F_{pool}W_2, \tag{5}$$

where $F_{pool} \in \mathbb{R}^{N_s \times c}$ denotes the pooling features, $\text{SPPool}(\cdot)$ represents the superpoint pooling operation, $F_v \in \mathbb{R}^{N_s \times d}$ indicates the visual features, $F_{sp} \in \mathbb{R}^{N_s \times d}$ denotes the superpoint features, and $W_1, W_2 \in \mathbb{R}^{c \times d}$ are both randomly initialized learnable parameters.

For a given text $T$, special tokens are added to both the beginning and the end of the text before it is input into the MPNet network (Song et al. 2020) to obtain token features. Given that this task requires the model to be capable of segmenting tokens at specified positions, we use the token features to generate the initial query $Q_0$:

$$Q_0 = EW_3 \tag{6}$$

where $E \in \mathbb{R}^{(L+2) \times e}$ represents the token features, and $W_3 \in \mathbb{R}^{e \times d}$ denotes the learnable parameters. Subsequently, $Q_0$ is fed into the decoder, where it first uses cross-attention (Vaswani et al. 2017) to integrate information from the visual modality, then employs self-attention to focus on the internal information within the sentence, and finally passes through a feed-forward network for nonlinear transformation. Additionally, we adopt a multi-layer architecture in series, meaning there are multiple such Cross-Self-FFN structures. For generality, the formula here is given for the $i$-th layer as an example:

$$Q_i = \text{FFN}(\text{Self}(\text{Cross}(Q_{i-1}, F_v))), i \in \{1, \ldots, N_l\}, \tag{7}$$

where $Q_{i-1}$ represents the query output from the previous layer and $N_l$ denotes the number of model layers. Finally, we compute the affinity between the query output from the last layer and the superpoint features, and then binarize this affinity to obtain the superpoint mask corresponding to the query. The superpoint mask can be broadcast to obtain point-level masks. For the sentence-level segmentation, use the mask corresponding to the [CLS] token as the result.

During the model training phase, we compute the superpoint mask for the query output from each layer and then calculate the loss against the ground truth. Here, we use the classic BCE loss and Dice loss (Milletari, Navab, and Ahmadi 2016). Additionally, similar to previous works (Wu et al. 2024b; Chen et al. 2025), we add an auxiliary Score loss. The final loss is formulated as follows:

$$L_{total} = \sum_{i=0}^{N_l} \lambda_1 L_{BCE}^i + \lambda_2 L_{Dice}^i + \lambda_3 L_{Score}^i \tag{8}$$

where the superscript $i$ denotes the layer number, and $\lambda_1$, $\lambda_2$, $\lambda_3$ are hyperparameters. $L_{total}$, $L_{BCE}$, $L_{Dice}$, $L_{Score}$ represent the total loss, BCE loss, Dice loss, and Score loss, respectively. During the model inference phase, only the output from the last layer is used as the final result.

# 5 Experiments

## 5.1 Experiment Settings

We utilize a Sparse 3D U-Net (Graham, Engelcke, and Van Der Maaten 2018) for extracting features from point clouds and a pre-trained MPNet (Song et al. 2020) for text feature extraction. The initial learning rate is set to 0.0001, which we decay at epoch 26, 34, 42 with a decay rate of 0.5. The number of model layers $N_l$ is set to 6. In the hyperparameters, $\lambda_1$, $\lambda_2$, and $\lambda_3$ are set to 1, 1, and 0.5, respectively.

| Method | Long | | | | Complex | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.25 | 0.5 | mIoU-S | mIoU | 0.25 | 0.5 | mIoU-S | mIoU | 0.25 | 0.5 | mIoU-S | mIoU |
| Val | | | | | | | | | | | | |
| PNG (González et al. 2021) | 50.3 | 34.0 | 35.8 | 35.1 | 52.7 | 35.4 | 37.0 | 36.4 | 57.2 | 41.9 | 42.6 | 41.3 |
| 3D-STMN (Wu et al. 2024b) | 63.8 | 46.0 | 46.4 | 45.2 | 65.4 | 48.8 | 48.0 | 47.1 | 71.8 | 55.8 | 53.8 | 52.7 |
| DetailBase (Ours) | **67.3** | **49.0** | **50.2** | **48.9** | **70.3** | **52.0** | **52.5** | **51.3** | **73.9** | **58.4** | **56.3** | **55.4** |
| Test | | | | | | | | | | | | |
| PNG (González et al. 2021) | 52.0 | 35.6 | 37.9 | 36.5 | 53.8 | 37.3 | 38.5 | 37.7 | 56.3 | 40.7 | 41.1 | 40.4 |
| 3D-STMN (Wu et al. 2024b) | 67.1 | 48.9 | 48.7 | 47.6 | 69.9 | 51.8 | 51.1 | 50.0 | 71.9 | 54.8 | 53.1 | 52.5 |
| DetailBase (Ours) | **71.1** | **52.8** | **52.5** | **51.5** | **73.5** | **55.8** | **54.8** | **53.8** | **74.8** | **58.5** | **56.2** | **55.7** |

Table 2: 3D-DRES results on DetailRefer dataset, showing mIoU, mIoU-S and accuracy at IoU thresholds of 0.25 and 0.5. "Long" indicates samples exceeding 50 tokens, while "Complex" refers to samples with four or more noun phrases to segment.

The batch size is set to 16. All experiments are trained on an NVIDIA GeForce RTX 3090 GPU.

## 5.2 Quantitative Results

Given that there currently exists no model that can directly adapt to the 3D-DRES task, we made appropriate adjustments to two existing models, namely PNG (González et al. 2021) and 3D-STMN (Wu et al. 2024b), to fit our task. The PNG model originates from the 2D domain's Panoptic Narrative Grounding task, which is quite similar to our task. We first changed its input by using point clouds instead of images, then utilized SPFormer (Sun et al. 2023) to extract instance proposals and averaged the point features within each instance as the instance feature. We also modified its result matching method, changing from selecting the candidate with the highest matching similarity to candidates with matching similarity above a certain threshold. This adjustment was made due to cases in our task where one noun phrase corresponds to multiple instances. For 3D-STMN, we altered its supervision method and result generation approach to align with DetailBase. Additionally, in the original 3D-STMN, the 3D-UNet was frozen. To ensure fairness, we have opted to include it in the training process here.

In Tab. 2, we present the performance of each model on both the validation set and the test set. It can be observed that although the modified PNG model achieved an miou of 40.4 on the test set, it still significantly lags behind other models. The reason for this situation is that the current effectiveness of 3D instance segmentation networks is not ideal; such two-stage models generally perform weaker than one-stage models. The modified 3D-STMN demonstrated its capabilities in this task, but due to the low compatibility of its designed modules with the task, its performance fell behind our DetailBase. DetailBase achieved an miou of 55.7 on the test set, a result that lays a suitable foundation for the further development of 3D-DRES methods. Moreover, our model framework is relatively simple and highly scalable, making it very appropriate as an initial approach for this task.

Phrase-level segmentation emphasizes fine-grained semantic understanding, whereas sentence-level segmentation focuses more on holistic comprehension. These two are not mutually exclusive; instead, they are mutually beneficial. We conducted joint training experiments to substantiate this viewpoint. By treating the [CLS] token (the root node in 3D-STMN) as the "noun phrase" to be segmented for 3D-RES tasks, we unified the formats of both tasks. We performed separate training on ScanRefer (Chen, Chang, and Nießner 2020) and DetailRefer datasets, as well as joint training across both datasets, reporting the mIoU of the models on the validation set in the Tab. 3. It is evident that joint training yields superior results compared to separate training for both 3D-STMN and DetailBase. Notably, joint training significantly enhances performance on 3D-RES, improving scores by 2.8 points on DetailBase and up to 3.2 points on 3D-STMN. In conclusion, our task not only holds its unique value but also complements traditional tasks synergistically.

| **Dataset** | **DetailBase** | | **3D-STMN**(Wu et al. 2024b) | |
|---|---|---|---|---|
| | 3D-RES | 3D-DRES | 3D-RES | 3D-DRES |
| Scanrefer | 44.0 | - | 41.8 | - |
| DetailRefer | - | 55.4 | - | 52.7 |
| Both | 46.8 | 56.8 | 45.0 | 53.3 |

Table 3: Results of separate training and joint training.

| $N_l$ | **Long** | **Complex** | **Overall** | | |
|---|---|---|---|---|---|
| | | | 0.25 | 0.5 | mIoU |
| 1 | 46.7 | 49.0 | 74.4 | 54.1 | 53.1 |
| 3 | 48.6 | 50.8 | 73.8 | 58.2 | 55.1 |
| 6 | **48.9** | 51.3 | **73.9** | 58.4 | 55.4 |
| 9 | 48.2 | **51.4** | 71.8 | **59.7** | **55.5** |

Table 4: Ablation study on number of model layers.

## 5.3 Ablation Study

Ablation experiments are all conducted on the validation set.

We conducted an ablation study on the parameter $N_l$, which indicates the number of layers in the model. We

| **Multi_layer** | **Score** | **Long** | **Complex** | **Overall** |
|---|---|---|---|---|
| × | × | 45.0 | 46.3 | 50.5 |
| × | ✓ | 44.5 | 46.5 | 50.8 |
| ✓ | × | 48.4 | 50.9 | 55.0 |
| ✓ | ✓ | **48.9** | **51.3** | **55.4** |

Table 5: Ablation on loss. 'Multi_layer' indicates that each layer is supervised. 'Score' refers to auxiliary score loss.

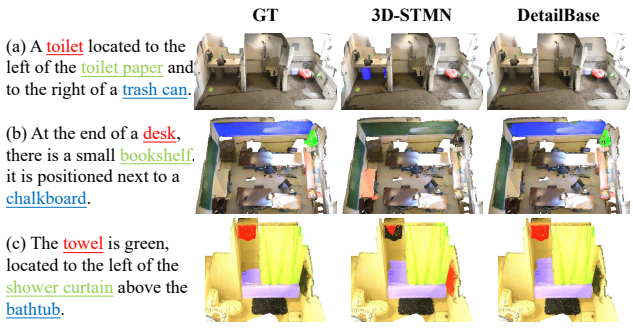


Figure 5: Comparison of visualization results. The 3D-STMN results are the model predictions after adaptation.


**Description:** There is a square chair. It is at a round table near a plant.

Figure 6: Visual performance of 3D-STMN on the ScanRefer dataset under different training methods.

present the mIoU for the 'long' and 'complex' subset, as well as the Acc@0.25, Acc@0.5, and mIoU on the entire validation set in Tab. 4. It can be observed that with only one layer, due to having fewer parameters and insufficient fitting capability, the model performed poorly, achieving an mIoU of only 53.1. When the number of layers reached three, there was a noticeable improvement in performance, with the mIoU increasing to 55.1. Beyond this point, the performance gains from increasing the number of layers diminished. Taking into account both model performance and complexity, six layers is the optimal choice.

We conducted an ablation study on the model's loss and presented the mIoU in Tab. 5. In this context, "Multi_layer" indicates that each layer is supervised; if not used, supervision only occurs at the final layer. "Score" refers to Score loss, which employs an MLP to predict the IoU. As shown in the table, applying supervision at every layer has a significant impact, improving the mIoU on the overall validation set by nearly five points. Regarding the Score loss, although its enhancement is relatively minor, as an auxiliary los, its computational cost is negligible. Therefore, adding the Score loss is a beneficial and harmless operation.

### 5.4 Qualitative Result

We present some visualization results in Fig. 5 to enable a more intuitive perception of the advantages of the 3D-DRES. As shown in (a), if it were a traditional 3D-RES, the text would refer to the object "toilet" and both models successfully segmented the target. It is difficult to assess a model's comprehension of the entire text at a fine-grained level, as the results are only evaluated based on a single entity within the text. However, under our 3D-DRES setting, we can observe the model's overall understanding of the text in a more detailed manner. For example, 3D-STMN demonstrates a completely incorrect understanding of text (b), whereas its comprehension of text (c) is partially correct. Through detailed analysis of model capabilities, we can better define the optimization directions for the model.

Additionally, we visualized the 3D-RES capabilities of 3D-STMN (Wu et al. 2024b) under different training methods, with one result shown in the Fig. 6 (more results in the appendix). It can be observed that when trained solely on the 3D-RES task, the model's fine-grained text understanding is poor. However, under joint training, the model's ability to capture fine-grained information within sentences significantly improves, enabling it to accurately locate auxiliary objects and thereby pinpoint the correct target.

## 6 Analysis of 3D-DRES

In this section, we briefly analyze the 3D-DRES task and summarize its challenges. A more detailed analysis, along with visual examples, is provided in Sec. 3 of the Appendix.

In our task, since the nouns to be segmented are specified and the variety of 3D scene objects is relatively small, the set of source noun phrases for generating segmentation kernels is not large. Additionally, approximately 10% of the noun phrases in our dataset correspond to multiple target objects. These two factors contribute to the model's difficulty in distinguishing between instances, especially when two objects of the same category are in close proximity. Distinguishing instances and determining the exact number of instances to be segmented present a significant challenge in this task.

Regarding the DetailBase framework, there is a category of text that poses significant challenges: texts involving instance-level clues. This is because the single-stage framework operates only at the superpoint level and lacks the ability to perceive information at the instance level. Additionally, the number of superpoints themselves limits their interactions, causing each superpoint to act almost as an independent entity. How to effectively utilize instance-level cues is also a major challenge in this task.

Finally, long texts are a distinctive feature of our dataset and also pose a challenge. Long texts imply more complex sentence structures, requiring the model to have a stronger ability to understand context.

## 7 Conclusion

In this paper, we introduce a novel task, 3D-DRES, which requires models to segment all noun phrases mentioned in sentences into corresponding masks. To support this task, we have constructed a new dataset, DetailRefer, featuring fine-grained annotations, combining both human effort and LLM. As the proposers of this task, we provide a highly scalable framework as a baseline for future researchers.

## Acknowledgements

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603, No.62525605), the National Natural Science Foundation of China (No. U21B2037 , No. 62302411) and China Postdoctoral Science Foundation (No. 2023M732948).

## References


Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*.

Azuma, D.; Miyanishi, T.; Kurita, S.; and Kawanabe, M. 2022. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*.

Chen, Q.; Wu, C.; Ji, J.; Ma, Y.; Yang, D.; and Sun, X. 2025. IPDN: Image-enhanced Prompt Decoding Network for 3D Referring Expression Segmentation. *arXiv:2501.04995*.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022. Language conditioned spatial relation reasoning for 3d object grounding. *NeurIPS*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*.

Chng, Y. X.; Zheng, H.; Han, Y.; Qiu, X.; and Huang, G. 2024. Mask grounding for referring image segmentation. In *CVPR*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*.

Dang, J.; Chen, L.; Wu, J.; Lin, R.; Wang, B.; Wang, Y.; Wang, L.; Zhu, N.; and Wang, T. 2025a. Diff-LMM: Diffusion Teacher-Guided Spatio-Temporal Perception for Video Large Multimodal Models.

Dang, J.; Deng, S.; Chang, H.; Wang, T.; Wang, B.; Wang, S.; Zhu, N.; Niu, G.; Zhao, J.; and Liu, J. 2025b. Hallucination Reduction in Video-Language Models via Hierarchical Multimodal Consistency.

Ding, Z.; Ding, Z.-h.; Hui, T.; Huang, J.; Wei, X.; Wei, X.; and Liu, S. 2022. Ppmn: Pixel-phrase matching network for one-stage panoptic narrative grounding. In *ACM MM*.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. Free-form description guided 3d visual graph network for object grounding in point cloud. In *ICCV*.

González, C.; Ayobi, N.; Hernández, I.; Hernández, J.; Pont-Tuset, J.; and Arbeláez, P. 2021. Panoptic narrative grounding. In *ICCV*.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*.

Guo, W.; Xu, X.; Wang, Z.; Feng, J.; Zhou, J.; and Lu, J. 2025. Text-guided Sparse Voxel Pruning for Efficient 3D Visual Grounding. *arXiv:2502.10392*.

He, S.; and Ding, H. 2024. RefMask3D: Language-guided transformer for 3D referring segmentation. In *ACM MM*.

He, S.; Ding, H.; Jiang, X.; and Wen, B. 2024. Segpoint: Segment any point cloud via large language model. In *ECCV*.

He, S.; Ji, P.; Yang, Y.; Wang, C.; Ji, J.; Wang, Y.; and Ding, H. 2025a. A survey on 3d gaussian splatting applications: Segmentation, editing, and generation. *arXiv:2508.09977*.

He, S.; Jie, G.; Wang, C.; Zhou, Y.; Hu, S.; Li, G.; and Ding, H. 2025b. ReferSplat: Referring segmentation in 3d gaussian splatting. *arXiv:2508.08252*.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *ECCV*.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2021. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*.

Hui, T.; Ding, Z.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Enriching phrases with coupled pixel and object contexts for panoptic narrative grounding. *arXiv:2311.01091*.

Jain, A.; Gkanatsios, N.; Mediratta, I.; and Fragkiadaki, K. 2022. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *CVPR*.

Landrieu, L.; and Simonovsky, M. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*.

Li, H.; Hui, T.; Ding, Z.; Zhang, J.; Ma, B.; Wei, X.; Han, J.; and Liu, S. 2024a. Dynamic prompting of frozen text-to-image diffusion models for panoptic narrative grounding. In *ACM MM*.

Li, R.; Li, S.; Kong, L.; Yang, X.; and Liang, J. 2024b. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. *arXiv:2412.04383*.

Liu, C.; Ding, H.; and Jiang, X. 2023. Gres: Generalized referring expression segmentation. In *CVPR*.

Liu, X.; Xu, X.; Li, J.; Zhang, Q.; Wang, X.; Sebe, N.; and Ma, L. 2024. LESS: Label-Efficient and Single-Stage Referring 3D Segmentation. *arXiv:2410.13294*.

Luo, J.; Fu, J.; Kong, X.; Gao, C.; Ren, H.; Shen, H.; Xia, H.; and Liu, S. 2022. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *CVPR*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.

Mi, B.; Wang, H.; Wang, T.; Chen, Y.; and Pang, J. 2025. Evolving Symbolic 3D Visual Grounder with Weakly Supervised Reflection. *arXiv:2502.01401*.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*.

Peng, Q.; Zheng, H.; and Huang, G. 2025. ProxyTransformation: Preshaping Point Cloud Manifold With Proxy Attention For 3D Visual Grounding. *arXiv:2502.19247*.

Pham, K.; Huynh, C.; Lim, S.-N.; and Shrivastava, A. 2024. Composing object relations and attributes for image-text matching. In *CVPR*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Qian, Z.; Ma, Y.; Lin, Z.; Ji, J.; Zheng, X.; Sun, X.; and Ji, R. 2024. Multi-branch Collaborative Learning Network for 3D Visual Grounding. In *ECCV*.

Shen, H.; Zhao, T.; Zhu, M.; and Yin, J. 2024. Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In *AAAI*.

Shi, X.; Wu, Z.; and Lee, S. 2024. Aware Visual Grounding in 3D Scenes. In *CVPR*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *NeurIPS*.

Sun, J.; Qing, C.; Tan, J.; and Xu, X. 2023. Superpoint transformer for 3d scene instance segmentation. In *AAAI*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*.

Wan, D.; Cho, J.; Stengel-Eskin, E.; and Bansal, M. 2024. Contrastive region guidance: Improving grounding in vision-language models without training. In *ECCV*.

Wang, H.; Ji, J.; Zhou, Y.; Wu, Y.; and Sun, X. 2023a. Towards real-time panoptic narrative grounding by an end-to-end grounding network. In *AAAI*.

Wang, X.; Zhao, N.; Han, Z.; Guo, D.; and Yang, X. 2025a. AugRefer: Advancing 3D Visual Grounding via Cross-Modal Augmentation and Spatial Relation-based Referring. *arXiv:2501.09428*.

Wang, Y.; Ding, H.; He, S.; Jiang, X.; Wei, B.; and Liu, J. 2025b. Hierarchical Alignment-enhanced Adaptive Grounding Network for Generalized Referring Expression Comprehension. *arXiv:2501.01416*.

Wang, Y.; Ni, J.; Liu, Y.; Yuan, C.; and Tang, Y. 2025c. IteRPrimE: Zero-shot Referring Image Segmentation with Iterative Grad-CAM Refinement and Primary Word Emphasis. *arXiv:2503.00936*.

Wang, Z.; Huang, H.; Zhao, Y.; Li, L.; Cheng, X.; Zhu, Y.; Yin, A.; and Zhao, Z. 2023b. 3drp-net: 3d relative position-aware network for 3d visual grounding. *arXiv:2307.13363*.

Wu, C.; Liu, Y.; Ji, J.; Ma, Y.; Wang, H.; Luo, G.; Ding, H.; Sun, X.; and Ji, R. 2024a. 3d-gres: Generalized 3d referring expression segmentation. In *ACM MM*.

Wu, C.; Ma, Y.; Chen, Q.; Wang, H.; Luo, G.; Ji, J.; and Sun, X. 2024b. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *AAAI*.

Wu, Y.; Cheng, X.; Zhang, R.; Cheng, Z.; and Zhang, J. 2023. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*.

Wu, Z.; Li, H.; Chen, G.; Yu, Z.; Gu, X.; and Wang, Y. 2024c. 3d question answering with scene graph reasoning. In *ACM MM*.

Xu, W.; Shi, C.; Tu, S.; Zhou, X.; Liang, D.; and Bai, X. 2025. A unified framework for 3d scene understanding. *NeurIPS*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2. 5 technical report. *arXiv:2412.15115*.

Yang, D.; Ji, J.; Ma, Y.; Guo, T.; Wang, H.; Sun, X.; and Ji, R. 2024b. Sam as the guide: mastering pseudo-label refinement in semi-supervised referring expression segmentation. *arXiv:2406.01451*.

Yang, D.; Ji, J.; Sun, X.; Wang, H.; Li, Y.; Ma, Y.; and Ji, R. 2023a. Semi-supervised panoptic narrative grounding. In *ACM MM*.

Yang, L.; Zhang, Z.; Qi, Z.; Xu, Y.; Liu, W.; Shan, Y.; Li, B.; Yang, W.; Li, P.; Wang, Y.; et al. 2023b. Exploiting contextual objects and relations for 3d visual grounding. *NeurIPS*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*.

Yang, Z.; Zhang, S.; Wang, L.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*.

Yuan, Z.; Yan, X.; Liao, Y.; Guo, Y.; Li, G.; Cui, S.; and Li, Z. 2022. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *CVPR*.

Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Wang, S.; Li, Z.; and Cui, S. 2021. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *ICCV*.

Zhan, Y.; Zhu, Y.; Chen, Z.; Yang, F.; Tang, M.; and Wang, J. 2024. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*.

Zhang, H.; Li, H.; Li, F.; Ren, T.; Zou, X.; Liu, S.; Huang, S.; Gao, J.; Leizhang; Li, C.; et al. 2024. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*.

Zhang, Y.; Gong, Z.; and Chang, A. X. 2023. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*.

Zhu, C.; Wang, T.; Zhang, W.; Chen, K.; and Liu, X. 2024. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *ECCV*.