DDPA-3DVG: Vision-Language Dual-Decoupling and Progressive Alignment for 3D Visual Grounding

Hongjie Gu[1], **Jinlong Fan**[1*], **Liang Zheng**[1], **Jing Zhang**[2], **Yuxiang Yang**[1*]
[1]School of Electronics and Information, Hangzhou Dianzi University, China
[2]School of Computer Science, Wuhan University, China
{hongjiegu, jfan, zhlbsbx, yyx}@hdu.edu.cn, jingzhang.cv@gmail.com

Abstract

3D visual grounding aims to localize target objects in point clouds based on free-form natural language, which often describes both target and reference objects. Effective alignment between visual and text features is crucial for this task. However, existing two-stage methods that rely solely on object-level features can yield suboptimal accuracy, while one-stage methods that align only point-level features can be prone to noise. In this paper, we propose DDPA-3DVG, a novel framework that progressively aligns visual locations and language descriptions at multiple granularities. Specifically, we decouple natural language descriptions into distinct representations of target objects, reference objects, and their mutual relationships, while disentangling 3D scenes into object-level, voxel-level, and point-level features. By progressively fusing these dual-decoupled features from coarse to fine, our method enhances cross-modal alignment and achieves state-of-the-art performance on three challenging benchmarks—ScanRefer, Nr3D, and Sr3D. The code will be released at https://github.com/HDU-VRLab/DDPA-3DVG.

1 Introduction

The goal of Visual Grounding (VG) [Deng *et al.*, 2021; Liu *et al.*, 2024] is to precisely locate the object referred to by a natural language description. In recent years, the progress of 3D data acquisition devices has sparked growing interest in using point clouds for various vision and language tasks, such as Dense Captioning (DC) [Yuan *et al.*, 2022; Wang *et al.*, 2022] and Visual Grounding (VG) [Yang *et al.*, 2020; Li and Sigal, 2021]. Among these, 3D visual grounding (3D VG) plays a crucial role in environmental perception, ranging from visual language navigation to autonomous robotics [Chen *et al.*, 2021; Chen *et al.*, 2022; Gao *et al.*, 2023; Liu *et al.*, 2023; Wang *et al.*, 2023a]. However, bridging the gap between the differing modalities of natural language and 3D scenes remains challenging.

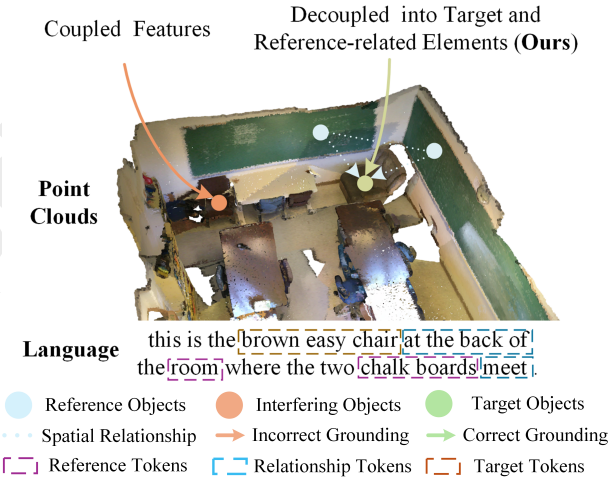

Figure 1: In complex scenarios, coupled sentence-level and scene-level features introduce ambiguity in 3D grounding. Our method decouples language/visual features into target/reference elements for precise grounding.

Existing methods primarily follow two paradigms: the two-stage paradigm [Zhao *et al.*, 2021; Yang *et al.*, 2021; Zhang *et al.*, 2023; Guo *et al.*, 2023; Yuan *et al.*, 2021] and the one-stage paradigm [Luo *et al.*, 2022; Jain *et al.*, 2022; Wu *et al.*, 2023]. In the two-stage paradigm, a universal 3D detector first generates numerous candidates. These proposals interact extensively with the text features to produce language-modulated visual representations, which are then scored by a regression head to yield the final prediction. Conversely, the one-stage paradigm typically uses Farthest Point Sampling (FPS) to select a fixed set of points. Under the guidance of text features, the points with the highest confidence are progressively identified to achieve grounding.

Accurate alignment between visual representations and language descriptions is central to the success of 3D VG. Nonetheless, current approaches either keep the information in natural language and 3D scenes fully entangled or, conversely, fragment it too finely for alignment—both resulting in suboptimal performance. For example, in two-stage approaches, only object-level features are utilized, making fine-grained target object localization difficult. In contrast,

*Corresponding author.

one-stage approaches rely exclusively on point-level features, which can be too noisy to capture precise semantic cues and often produce diffuse confidence distributions. A similar challenge occurs with natural language: if treated solely at the sentence level, crucial distinctions between target and reference objects may be masked. While recent work [Wu *et al.*, 2023] partially addresses imbalance and ambiguity by explicitly modeling language as a dependency tree, *the problem of effectively extracting meaningful guidance from natural language and representing visual features at an appropriate granularity for cross-modal alignment remains unsolved.*

In this paper, we propose a novel method, dubbed DDPA-3DVG, that transforms 3D scene information into multi-grained visual features under text guidance, while also decoupling natural language descriptions into coherent semantic components to improve target localization. We then align these dual-decoupled features from coarse to fine, achieving more precise 3D visual grounding. As illustrated in Figure 1, when dealing with complex 3D scenes containing multiple similar target and reference objects, existing methods that rely on fully coupled features often suffer from ambiguous alignments. By decoupling both the scene and language into more focused elements, our approach makes the alignment process more systematic and precise, especially when fusing text-guided localization with multi-granularity visual representations. Specifically, we divide the 3D scene into object-level, voxel-level, and point-level features, while splitting natural language into target object-related features and reference object-related features. These dual-decoupled features are then gradually fused from coarse to fine, allowing the network to effectively match target and reference descriptions with their corresponding visual features. Moreover, to effectively capture both global and local context, we employ static queries distributed across the entire scene alongside position-learnable dynamic queries. During the alignment process, these mixed queries interact progressively with the dual-decoupled features, facilitating a more comprehensive understanding of 3D spatial information and linguistic cues while minimizing interference from irrelevant objects.

In summary, our main contributions are:

- **Dual-Decoupling Strategy:** We decouple both the 3D scene and natural language to extract multi-granularity visual features and semantically rich text features, facilitating more effective cross-modal alignment.
- **Progressive Alignment Framework:** We progressively fuse these dual-decoupled features in a coarse-to-fine manner, which significantly improves localization precision for the target object.
- Experiments on three challenging benchmarks, ScanRefer, Nr3D, and Sr3D, demonstrate that our method achieves state-of-the-art performance, validating its effectiveness in 3D visual grounding.

2 Related Work

2.1 3D Vision-Language Tasks

Vision and language are essential modalities for understanding and interacting with the world. In particular, 3D Vision-Language tasks [Luo *et al.*, 2022; Azuma *et al.*, 2022] integrate 3D visual perception with linguistic information, enabling more comprehensive analyses of complex environments. One representative example is 3D Visual Question Answering (3D-VQA) [Ma *et al.*, 2022; Azuma *et al.*, 2022], which relies on point clouds as input and requires models to interpret object attributes and spatial relationships in order to provide contextually accurate answers. Another example is 3D Dense Captioning (3D DC) [Yuan *et al.*, 2022; Wang *et al.*, 2022], aimed at generating text descriptions for each object by identifying its position, attributes, and spatial relations within the scene. By contrast, 3D Visual Grounding [Luo *et al.*, 2022; Wu *et al.*, 2023] specifically targets localizing a particular object based on a given linguistic expression. In this work, we concentrate on advancing the model's ability to integrate multi-grained 3D vision representations and decoupled linguistic elements, thereby attaining more precise grounding performance.

2.2 3D Visual Grounding

3D visual grounding focuses on locating objects within a 3D scene based on free-form natural language descriptions. Two main paradigms have emerged in this area: two-stage methods [Yang *et al.*, 2021; Guo *et al.*, 2023; Yuan *et al.*, 2021] and one-stage methods [Luo *et al.*, 2022; Jain *et al.*, 2022; Wu *et al.*, 2023]. Two-stage approaches leverage off-the-shelf, pre-trained object detectors to generate object candidates, then match these proposals to linguistic features to select the target region. For example, 3DVG-Transformer [Zhao *et al.*, 2021] adopts a Transformer-inspired architecture to incorporate contextual cues, thereby enhancing proposal generation and cross-modal disambiguation. Multi3DRefer [Zhang *et al.*, 2023] extends the ScanRefer [Chen *et al.*, 2020] dataset and task to support multiple object references, introducing new evaluation metrics and benchmarks to accommodate more general grounding scenarios.

In contrast, one-stage approaches directly regress the bounding box of the target object, bypassing explicit proposal generation. 3D-SPS [Luo *et al.*, 2022] was among the first to refine localization guided by linguistic cues, while EDA [Wu *et al.*, 2023] relies on implicit text decoupling to separate key linguistic components for greater alignment precision. However, EDA solely decouples text features, leaving visual features unaddressed. In this work, we present DDPA-3DVG, which explicitly decouples both visual and text representations, enabling more effective cross-modal alignment for enhanced 3D grounding accuracy.

3 Method

3.1 Overview

An overview of our DDPA-3DVG framework is illustrated in Figure 2. First, we use a Vision Encoder and a Language Encoder to extract the visual and text features, respectively. A Cross-Modality Encoder then fuses these features (Sec. 3.2). Next, the language descriptions are split into semantically meaningful elements, generating text features for both the target-related component and reference-related component.

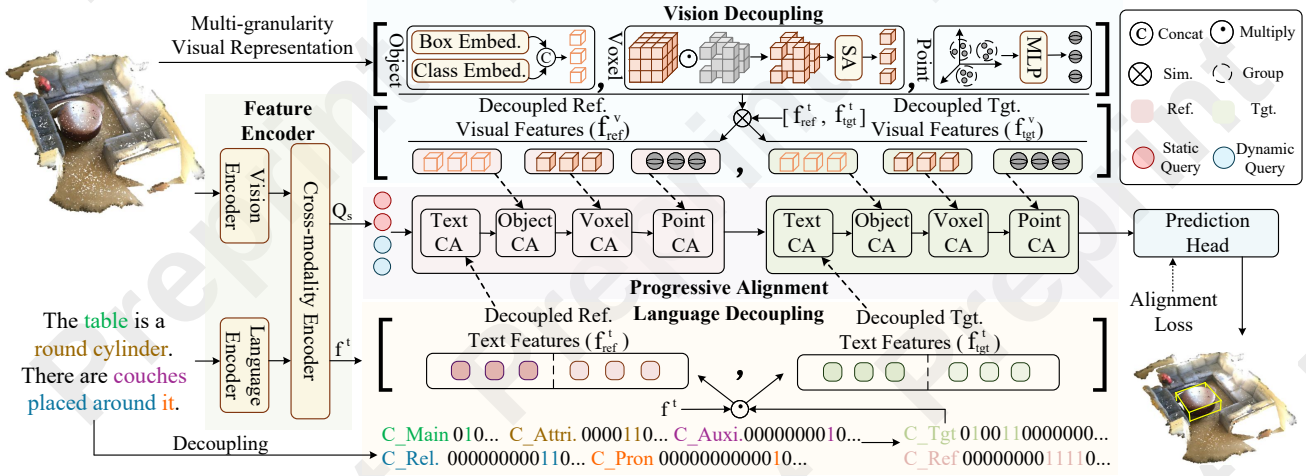

Figure 2: Overview of the Proposed DDPA-3DVG. Our method begins with the Feature Encoder module to jointly encode visual and linguistic information. Next, the Language Decoupling module splits text features into separate reference and target components. Guided by these decoupled text embeddings, the Vision Decoupling module then generates multi-grained visual features tailored to reference and target objects, respectively. In the Progressive Alignment module, both decoupled visual and text features are gradually fused with input queries, ultimately guiding the Prediction Head to accurately localize the target object. Two alignment losses are employed to optimize the decoupling and alignment processes.

Leveraging these decoupled text features, we decompose the 3D scene into multi-grained visual representations (Sec. 3.3). To further align the resulting decoupled visual and text features, we progressively fuse them with the input queries via cascaded reference and target decoders (Sec. 3.4). Finally, the fused queries localize the target object through a position alignment loss and a semantic alignment loss (Sec. 3.5).

3.2 Feature Encoder

For the language modality, we employ a pre-trained RoBERTa [Liu *et al.*, 2019] model to encode the input sentences. On the visual side, we generate point-level features $\mathbf{f}_p^v \in \mathbb{R}^{N \times D}$ through PointNet++ [Qi *et al.*, 2017], where N represents the number of points in the point cloud and D their feature dimensionality. Following [Jain *et al.*, 2022], we employ a Cross-Modality Encoder—consisting of multiple cross-attention layers—to integrate the extracted visual and text features. The output, $\mathbf{f}^t \in \mathbb{R}^{l \times D}$, where l is the length of the input sentence, is then used in the subsequent sentence decoupling stage; additionally, a set of static queries $Q_s \in \mathbb{R}^{N_s \times D}$ is initialized for the following progressive alignment phase.

3.3 Vision and Language Dual-decoupling

Language Decoupling

Drawing inspiration from EDA [Wu *et al.*, 2023], we use existing NLP tools [Schuster *et al.*, 2015; Wu *et al.*, 2019] to explicitly classify words in sentences into five categories:

1) *Main Object (Target):* The object to be localized. 2) *Attributes:* The target object's properties (e.g., color, shape). 3) *Auxiliary Objects (References):* Additional objects that help localize the target. 4) *Pronouns:* Referents for the main object. 5) *Relations:* Spatial relationships between the main object and auxiliary objects.

While EDA builds a dependency tree to decouple sentences, we further classify all words into two components based on these categories: the target component (composed of "main objects" and "attributes"), and the reference component (including "auxiliary objects", "pronouns", "relations" and any remaining ones). First, we encode the five word categories as multi-hot vectors, where the corresponding word position is marked as one. This process produces the codes $\{\mathbf{C}_{main}, \mathbf{C}_{attri}, \mathbf{C}_{auxi}, \mathbf{C}_{rel}, \mathbf{C}_{pron}, \mathbf{C}_{\varnothing}\}$, where $\mathbf{C}_{\varnothing}$ represents words that do not belong to any specific category. The target and reference components are then encoded as $\mathbf{C}_{tgt} = \mathbf{C}_{main} + \mathbf{C}_{attri}$ and $\mathbf{C}_{ref} = \mathbf{C}_{auxi} + \mathbf{C}_{rel} + \mathbf{C}_{pron} + \mathbf{C}_{\varnothing}$.

Next, we obtain the decoupled target and reference text features by applying the respective components to the text features $\mathbf{f}^t$:

$$\mathbf{f}^t_{\text{tgt}} = \mathbf{f}^t \cdot \mathbf{C}_{tgt}, \quad \mathbf{f}^t_{\text{ref}} = \mathbf{f}^t \cdot \mathbf{C}_{ref}. \tag{1}$$

This decomposition effectively separates the target object from the reference elements, enhancing the network's ability to interpret objects and their interrelationships within the scene.

Vision Decoupling

To capture both coarse and fine spatial cues, we propose multi-grained visual representations [Yuan *et al.*, 2025a; Yuan *et al.*, 2025b] at the object, voxel, and point levels. This approach benefits from high-level features for swift object identification and low-level features for precise localization.

Multi-granularity Visual Representations. Object-level features, commonly used in two-stage 3D VG approaches [Jain *et al.*, 2022], are extracted by applying a pre-trained 3D detector to the scene. The resulting bounding

boxes yield class and box embeddings, concatenated to form the object feature $\mathbf{f}^{vo}$.

Point-level features follow the one-stage 3D VG paradigm, where each point in the cloud is encoded by PointNet++, and FPS is applied to get point proposals. Local neighborhoods are then aggregated to refine each point feature $\mathbf{f}^{vp}$.

To bridge object- and point-level representations, we introduce voxel-level features. We divide the point cloud into a $W \times L \times H$ voxel grid:

$$i = \left\lfloor \frac{x}{\Delta W} \right\rfloor, \quad j = \left\lfloor \frac{y}{\Delta L} \right\rfloor, \quad k = \left\lfloor \frac{z}{\Delta H} \right\rfloor, \tag{2}$$

where (x, y, z) are a point's coordinates and (i, j, k) its voxel index. Each voxel's embedding $\bar{\mathbf{f}}^{vv}$ is taken as the average of its points' features. A shallow MLP calculates a usefulness score for each voxel, retaining only the top $R \times N_{\text{voxel}}$ based on this score:

$$\bar{\mathbf{f}}^{vm} = \text{TopK}\Big(\text{MLP}(\bar{\mathbf{f}}^{vv}),\ R \times N_{\text{voxel}}\Big). \tag{3}$$

Position embeddings are then added, followed by a self-attention layer for aggregating global context:

$$\mathbf{f}^{vv} = \text{SelfAttention}\big(\bar{\mathbf{f}}^{vv} \odot \bar{\mathbf{f}}^{vm} + \text{PE}(x_v)\big), \tag{4}$$

where x_v is the voxel's position, $\odot$ is element-wise multiplication. Finally, our multi-grained visual features are:

$$\mathbf{f}^{v} = \big[\mathbf{f}^{vo},\ \mathbf{f}^{vv},\ \mathbf{f}^{vp}\big]. \tag{5}$$

Text-guided Vision Decoupling. Since $\mathbf{f}^v$ represents the entire scene without text context, we further decouple it under the guidance of language description. We first compute the cosine similarity between each element in $\mathbf{f}^v$ and the target or reference text features $\mathbf{f}^t_{\text{tgt}}$ and $\mathbf{f}^t_{\text{ref}}$, then select the top-$K$ features with the highest similarity scores:

$$\mathbf{f}^v_{\text{tgt}},\ \mathbf{f}^v_{\text{ref}} = \text{TopK}\big(\mathbf{f}^v, \mathbf{f}^v \otimes \mathbf{f}^t_{\text{tgt}}\big),\ \text{TopK}\big(\mathbf{f}^v, \mathbf{f}^v \otimes \mathbf{f}^t_{\text{ref}}\big), \tag{6}$$

where $\otimes$ means the calculation of cosine similarity. The decoupled features thereafter become $\mathbf{f}^v_{\text{tgt}} = [\mathbf{f}^{vo}_{\text{tgt}}, \mathbf{f}^{vv}_{\text{tgt}}, \mathbf{f}^{vp}_{\text{tgt}}]$ and $\mathbf{f}^v_{\text{ref}} = [\mathbf{f}^{vo}_{\text{ref}}, \mathbf{f}^{vv}_{\text{ref}}, \mathbf{f}^{vp}_{\text{ref}}]$. This filtration discards distracting elements in the scene and improves alignment accuracy by focusing on the most salient features.

3.4 Progressive Alignment

After the dual-decoupling of vision and language, we obtain $\mathbf{f}^t_{\text{tgt}}, \mathbf{f}^t_{\text{ref}}, \mathbf{f}^v_{\text{tgt}}$, and $\mathbf{f}^v_{\text{ref}}$. We then propose a progressive alignment module composed of two cascaded decoders: 1) A *Reference Decoder* that progressively fuses $\mathbf{f}^t_{\text{ref}}$ with the multi-grained visual features $\mathbf{f}^v_{\text{ref}}$. 2) A *Target Decoder* that further refines $\mathbf{f}^t_{\text{tgt}}$ with $\mathbf{f}^v_{\text{tgt}}$.

Within each decoder, text features are merged with the object-, voxel-, and point-level features from coarse to fine. To capture both global and local spatial cues, we synthesize queries from two sources: 1) static queries derived from Farthest Point Sampling, initialized by Q_s, and 2) dynamic queries, learnable embeddings randomly initialized. During training, dynamic queries adaptively focus on crucial regions (e.g., the target object), while static queries help maintain broad scene coverage.

For example, consider a scenario where the text describes a reference object ("it is under a big painting that is gray, yellow, and brown") and a target object ("this is a light brown couch"). In the *Reference Decoder*, dynamic queries converge around the "painting" and obtain its spatial relationship "under" relative to the "couch." In the *Target Decoder*, the mentioned attributes of the couch guide the dynamic queries to accurately localize the correct furniture piece.

3.5 Prediction Head and Alignment Loss

After the progressive alignment process, the final queries are used to predict the target object's bounding box via a single-layer MLP and estimate the position labels, $\mathbf{C}_{pred}$, for each word through another MLP.

To enforce cohesive visual-text alignment, we define two losses: position alignment loss and semantic alignment loss. The position alignment loss is formulated as follows:

$$\mathcal{L}_{\text{pos}} = \sum_{i=1}^{k} \mathbf{C}^i_{\text{lang}} \log\Big(\frac{\mathbf{C}^i_{\text{lang}}}{\mathbf{C}^i_{\text{obj}}}\Big), \tag{7}$$

where the ground truth language distribution, $\mathbf{C}_{\text{lang}} \in \{\mathbf{C}_{\text{ref}}, \mathbf{C}_{\text{tgt}}\}$, is determined by the text position labels of each component. $\mathbf{C}_{\text{obj}}$ represents the semantic distributions of object features for the i-th candidate among k objects and is obtained by applying a softmax operation to $\mathbf{C}_{\text{pred}}$.

The semantic alignment loss measures coherence between object embeddings, $\mathbf{o}$, and text embeddings, $\mathbf{t}$. These embeddings are derived through linear projections of the output queries and $\mathbf{f}^t$, respectively. This loss is defined as:

$$\mathcal{L}^{\{F,N\}}_{\text{sem}} = \frac{1}{|\mathbf{F}^+_i|} \sum_{i=1}^{N} -\log\Big(\frac{\exp\big(\mathbf{s}^+_i/\tau\big)}{\exp\big(\mathbf{s}^+_i/\tau\big) + \exp\big(\mathbf{s}^-_i/\tau\big)}\Big), \tag{8}$$

where $\mathbf{s}^+_i$ and $\mathbf{s}^-_i$ denote the similarity scores between the object embedding $\mathbf{o}$ and the text feature vector $\mathbf{t}$ for matched and unmatched cases, respectively. F and N represent the positive features and the number of objects or text embeddings, respectively. The final semantic alignment loss is computed as $\mathcal{L}_{sem} = \frac{1}{2}(\mathcal{L}^{\{f^v,k\}}_{\text{sem}} + \mathcal{L}^{\{f^t,l\}}_{\text{sem}})$.

By minimizing these two losses, the model learns to align natural language descriptions with 3D scene features more effectively and encourages the visual and text representations of the same target to converge in feature space, thereby yielding improved accuracy on 3D visual grounding tasks.

4 Experiment

4.1 Datasets and Evaluation Metrics

We evaluated the effectiveness of DDPA-3DVG using three widely adopted and challenging datasets: ScanRefer [Chen *et al.*, 2020], Sr3D and Nr3D [Achlioptas *et al.*, 2020]. ScanRefer is a 3D visual grounding dataset constructed upon 800 scenes from ScanNet [Dai *et al.*, 2017], encompassing 51,583 text descriptions. Within each scene, if a target object is the only instance of its class, it is labeled as unique; alternatively, if there are multiple instances of that class, it is labeled as multiple. For ScanRefer, the evaluation uses Acc@mIoU,

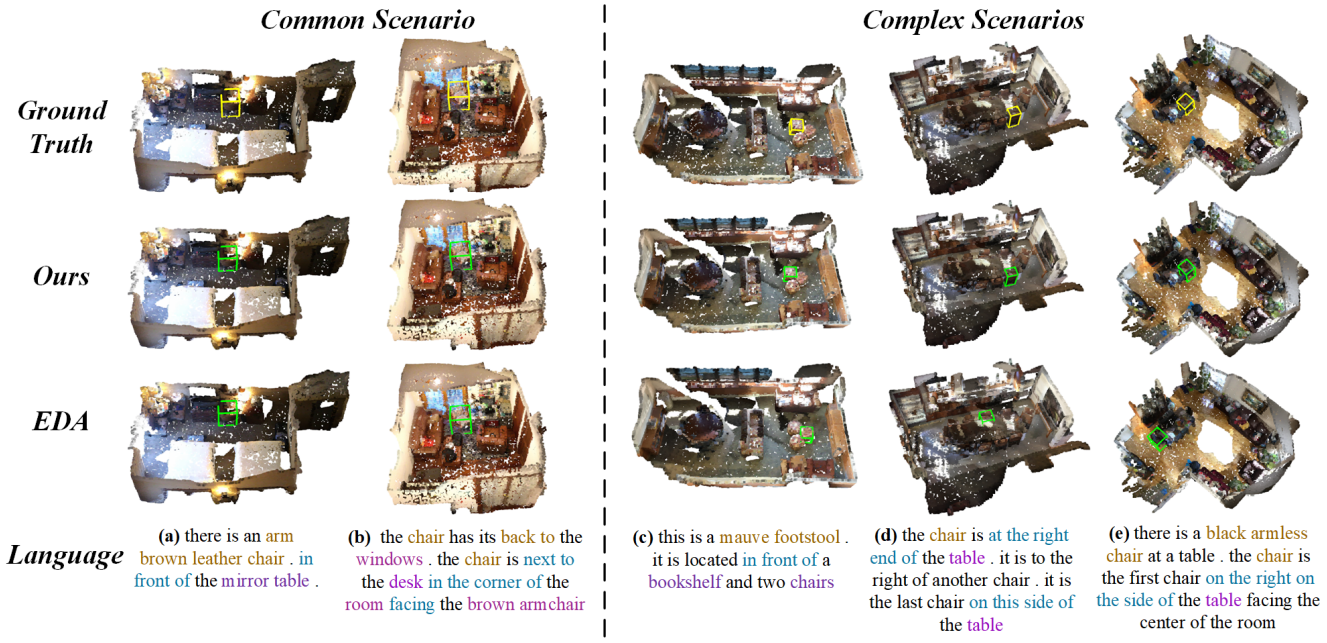

Figure 3: Qualitative results on the ScanRefer dataset. Visual results showcase both common and complex scenarios. Refer to the supplementary materials for additional details.

which measures the proportion of descriptions for which the predicted bounding box's Intersection over Union (IoU) with the ground truth exceeds m, where $m \in \{0.25, 0.5\}$. Similarly, Sr3D and Nr3D dataset [Achlioptas *et al.*, 2020], are also constructed from ScanNet. Nr3D contains 41,503 descriptions (annotated by humans) spanning 707 scenes, whereas Sr3D comprises 83,572 machine-generated descriptions. Both subsets use accuracy as the evaluation metric. Furthermore, each scene is categorized as easy or hard based on whether more than two instances of the same object class are present.

4.2 Comparison with State-of-the-Art Methods

Quantility and Visual Results

As presented in Table 1, we benchmarked our proposed DDPA-3DVG on the ScanRefer dataset against leading approaches. Our method achieves state-of-the-art accuracy rates of 55.3% and 43.2% for IoU thresholds 0.25 and 0.5, respectively, indicating robust performance. Notably, many earlier methods have leveraged auxiliary 2D image data to enhance point-cloud features [Chen *et al.*, 2020; Zhao *et al.*, 2021; Cai *et al.*, 2022], whereas our approach uses only sparse point clouds and still surpasses these image-dependent baselines. This result underscores the efficacy of the dual-decoupling mechanism and progressive alignment module we propose.

Following [Wu *et al.*, 2023], we further compared our approach with other one-stage methods (see the lower part of Table 1). Even without relying on object-level visual features, DDPA-3DVG outperforms competing alternatives, reinforcing the strength of the proposed design. We visualized our grounding results in Figure 3 for qualitative comparison. In typical scenes with fewer distractions, both EDA and our proposed method achieve high-precision grounding. However, in more complex scenarios where target identification is hindered by multiple interfering objects, our approach more effectively mitigates these distractors, consistently delivering correct predictions, which underscore our method's aptitude for understanding spatial contexts and accurately localizing targets in cluttered scenarios.

To confirm the generalizability of DDPA-3DVG, we also evaluated it on the SR3D and NR3D subsets of the ReferIt3D dataset [Jain *et al.*, 2022; Wu *et al.*, 2023]. As reported in Table 2, our method again outperforms other approaches, with accuracies of 54.4% on SR3D and 69.0% on NR3D. These results affirm the robustness of our approach across diverse 3D point-cloud environments.

To provide a more comprehensive look into the inner workings of DDPA-3DVG, we illustrated the progression of both static and dynamic queries in Figure 4 at four points during training: steps 0, 25, 50, and 100. Initially, the dynamic queries lie scattered throughout the 3D scene, showing no discernible structure. With successive training iterations, these queries begin to converge toward the reference object, precisely guided by the linguistic clues provided in the text descriptions. As the model refines its understanding of the spatial relationships and semantic attributes, the dynamic queries gradually adjust their positions from the reference zone toward the actual target region in the scene.

This gradual realignment process is driven by our dual-decoupling mechanism, which aims to isolate and refine relevant visual and text features. By decoupling vision and language at multiple levels, our method yields more discriminative representations, allowing the dynamic queries to interpret the scene context more precisely. The shift of the dy-

Methods	Modality	Unique(∼19%)		Multiple(∼81%)		**Overall**	
		0.25	0.5	0.25	0.5	**0.25**	**0.5**
3DJCG [Cai *et al.*, 2022]	3D+2D	83.5	64.3	41.4	30.8	49.6	37.3
BUTD-DETR [Jain *et al.*, 2022]	3D	82.9	65.0	44.7	34.0	50.4	38.6
EDA [Wu *et al.*, 2023]	3D	85.8	68.6	49.1	37.6	54.6	42.3
ViewRefer [Guo *et al.*, 2023]	3D	-	-	33.1	26.5	41.3	33.7
3DRP-Net [Wang *et al.*, 2023b]	3D	83.1	67.7	42.1	32.0	50.1	38.9
3DVLP [Zhang *et al.*, 2024]	3D	85.2	70.0	43.7	33.4	51.7	40.5
DDPA-3DVG	3D	86.8	70.2	49.8	38.4	55.3	43.3
3D-SPS* [Luo *et al.*, 2022]	3D	81.6	64.8	39.5	29.6	47.7	36.4
BUTD-DETR* [Jain *et al.*, 2022]	3D	81.5	61.2	44.2	32.8	49.8	37.1
EDA* [Wu *et al.*, 2023]	3D	86.4	69.4	48.1	36.8	53.8	41.7
DDPA-3DVG*	3D	86.5	69.9	48.6	37.4	54.3	42.2

Table 1: 3D visual grounding results on the ScanRefer dataset, with accuracy evaluated under IoU thresholds of 0.25 and 0.5. The upper section of the table reports the performance of two-stage methods, while the lower section presents results for single-stage methods. Single-stage methods (denoted with *) are implemented without the need for an additional 3D object detection step. For more comprehensive results, please refer to the supplementary materials.

Methods	Nr3D	Sr3D
BUTD-DETR [Jain *et al.*, 2022]	43.3	52.1
LAR [Bakr *et al.*, 2022]	48.9	59.4
3D-SPS [Luo *et al.*, 2022]	51.5	62.6
M3DRef-CLIP [Zhang *et al.*, 2023]	49.4	/
EDA [Wu *et al.*, 2023]	52.1	68.1
3DReftTR-SR [Lin *et al.*, 2023]	52.6	68.5
DDPA-3DVG	54.4	69.7

Table 2: Comparison of 3D visual grounding results with state-of-the-art methods on the SR3D and NR3D datasets. Performance is evaluated using accuracy under the 0.25 IoU metric. For more experimental results, please refer to the supplementary materials.

Methods	Subsets			**Overall**	
	A.	R.	A. + R.	**@0.25**	**@0.5**
ScanRefer	11.17	10.53	10.29	10.51	6.20
TGNN	10.52	13.32	11.35	11.64	9.51
InstanceRefer	14.74	13.71	13.81	13.92	11.47
BUTD-DETR	12.30	12.11	11.86	11.99	8.95
EDA	25.40	25.82	26.96	26.50	21.20
DDPA-3DVG	26.40	27.18	28.37	27.83	22.68

Table 3: Performance evaluation of 3D grounding on subsets without object names. Accuracy of subsets is reported based on the acc@0.25IoU metric. "A." and "R." represent attribution and relationship.

namic queries over time provides tangible evidence of how decoupled semantics and visual cues are integrated step by step. Critically, our progressive fusion strategy—carried out through specialized reference and target decoders—ensures that the model systematically narrows down potential targets, eventually pinpointing the correct object with higher accuracy. The consistent adjustment of query positions confirms that combining vision-language dual decoupling with the progressive alignment module is instrumental in tackling challenging 3D grounding tasks, especially within complex environments.

Grounding without Object Name

To assess our model's ability to reason beyond direct linguistic cues such as explicit object names, we adopted the "Grounding without Object Name" task proposed by EDA [Wu *et al.*, 2023], wherein all target names in the ScanRefer validation set are replaced with the token "object". EDA further divided the language descriptions into four subsets that focus on: (i) only attributes, (ii) only spatial relationships, and (iii) both attributes and relationships. As summarized in Table 3, our model, without additional training, achieves state-of-the-art results across all these subsets. This superior performance indicates that the reference object and spatial relationship information extracted via our decoupling framework has a highly effective guiding impact on the grounding process, enabling accurate localization even when explicit object names are omitted.

Results about Convergence

Figure 5 presents the experimental results. In terms of convergence, our method achieves faster convergence during training. Specifically, the number of epochs required for our method to reach a performance of 52.7% is 37 fewer than that of EDA, representing a 1.97× reduction in training epochs. Additionally, after convergence, our method achieves an average accuracy improvement of 1.23% compared to EDA when IoU is set to 0.25.

These results highlight the effectiveness of our dual decoupling of vision and language features and the progressive alignment mechanism, which enables the model to align cross-modal features more efficiently and effectively. This not only accelerates the training process but also enhances the overall model performance. In summary, our method outperforms EDA in terms of both convergence efficiency and effectiveness for 3D Visual Grounding.

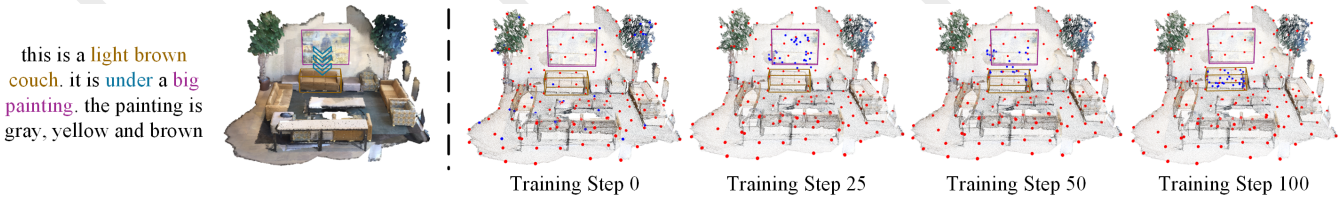

Figure 4: Illustration of the query convergence process during the learning phase. Target objects and reference objects are highlighted with bounding boxes in colors corresponding to the text. Red points represent static queries, while blue points represent dynamic queries.

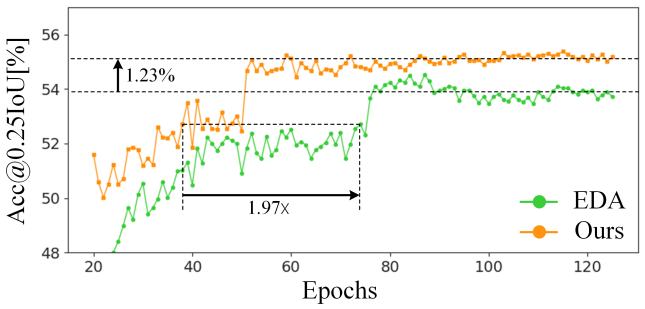

Figure 5: Comparison of convergence with EDA.

O	V	P	Unique(∼19%) 0.25	Unique(∼19%) 0.5	Multiple(∼81%) 0.25	Multiple(∼81%) 0.5	**Overall @0.25**	**Overall @0.5**
✓			71.9	51.63	35.7	24.3	43.8	33.1
		✓	72.4	51.1	36.3	24.8	44.6	34.0
✓		✓	85.5	69.1	48.3	37.7	54.1	41.8
✓	✓	✓	**86.8**	**70.2**	**49.8**	**38.4**	**55.3**	**43.3**

Table 4: Effectiveness of multi-granularity visual representations. O, V, and P denote object-level, voxel-level, and point-level representations, respectively.

4.3 Ablation Studies

Effectiveness of Mutil-granularity Visual Representation

To evaluate the effectiveness of multi-granularity visual representation, we trained four model variants under different representations, as reported in Table 4. Incorporating the object-, voxel-, and point-level features in a unified manner consistently enhances performance. This improvement arises from the enriched, hierarchically distinct 3D spatial information provided by voxel-level embeddings, which supplement the coarse-to-fine granularity of the visual features.

Effectiveness of Language Decoupling

To validate the effectiveness of language decoupling, we removed the decoupling process and instead used sentence-level features directly in the progressive alignment. When reference- and target-related linguistic cues are entangled, the decoder struggles to precisely align visual and linguistic features, resulting in noticeably inferior performance, as shown in Table 5. Specifically, accuracy decreases by 10.2% and 7.7%, respectively. This highlights the importance of separating target and reference-related elements, as the proposed feature decoupling provides fine-grained discriminative ability to guide precise visual grounding.

	Unique(∼19%) 0.25	Unique(∼19%) 0.5	Multiple(∼81%) 0.25	Multiple(∼81%) 0.5	**Overall @0.25**	**Overall @0.5**
w/o decoupling	86.4	68.8	48.9	37.2	54.5	42.3
DDPA-3DVG	**86.8**	**70.2**	**49.8**	**38.4**	**55.3**	**43.3**

Table 5: Removing language decoupling and using sentence-level features in progressive alignment results in inferior performance, demonstrating the importance of decoupling

Q_d	Q_s	Unique(∼19%) 0.25	Unique(∼19%) 0.5	Multiple(∼81%) 0.25	Multiple(∼81%) 0.5	**Overall @0.25**	**Overall @0.5**
✓		72.2	48.6	32.6	16.8	38.5	21.6
	✓	84.0	67.8	48.2	37.3	53.6	41.9
✓	✓	**86.8**	**70.2**	**49.8**	**38.4**	**55.3**	**43.3**

Table 6: Effectiveness of mixed input queries. Q_d denotes dynamic points, while Q_s denotes static points.

Effectiveness of Mixed Points

To investigate the benefits of mixed points in the decoder queries, we performed an ablation experiment with three configurations: (1) Q_s only, (2) Q_d only, and (3) the combination $Q_s + Q_d$. As detailed in Table 6, leveraging both static and dynamic point queries significantly improves localization accuracy. Dynamic points converge around the target object during training, whereas static points retain a broader receptive field and correct for instances that dynamic points might overlook. These two categories of points thus complement each other, resulting in more robust grounding performance.

5 Conclusion

In this paper, we present DDPA-3DVG, an effective approach for the 3D Visual Grounding task. We innovatively introduce a voxel-based, multi-granularity visual representation, endowing the model with hierarchical 3D spatial information. Furthermore, we decouple text and multi-granularity visual features into reference and target information using Language Decoupling and Text-guided Vision Decoupling modules. Our Progressive Alignment module then progressively grounds the target object, effectively mitigating interference from the 3D scene and notably improving localization precision. Though effective and demonstrating substantial improvements over existing methods, our approach could benefit from more efficient alignment modules and learnable candidates to eliminate dependence on pre-trained detectors, which remain valuable directions for future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376080), the Zhejiang Provincial Natural Science Foundation Key Fund of China (LZ23F030003), the National Natural Science Foundation of China (62476075), and the Zhejiang Key Laboratory of Optoelectronic Intelligent Imaging and Aerospace Sensing.

References

[Achlioptas *et al.*, 2020] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020.

[Azuma *et al.*, 2022] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022.

[Bakr *et al.*, 2022] Eslam Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. *Advances in Neural Information Processing Systems*, 35:37146–37158, 2022.

[Cai *et al.*, 2022] 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.

[Chen *et al.*, 2020] 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.

[Chen *et al.*, 2021] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:5834–5847, 2021.

[Chen *et al.*, 2022] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16537–16547, 2022.

[Dai *et al.*, 2017] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[Deng *et al.*, 2021] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[Gao *et al.*, 2023] Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. Adaptive zone-aware hierarchical planner for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14911–14920, 2023.

[Guo *et al.*, 2023] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding with gpt and prototype guidance. *arXiv preprint arXiv:2303.16894*, 2023.

[Jain *et al.*, 2022] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022.

[Li and Sigal, 2021] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34:19652–19664, 2021.

[Lin *et al.*, 2023] Haojia Lin, Yongdong Luo, Xiawu Zheng, Lijiang Li, Fei Chao, Taisong Jin, Donghao Luo, Yan Wang, Liujuan Cao, and Rongrong Ji. A unified framework for 3d point cloud visual grounding. *arXiv preprint arXiv:2308.11887*, 2023.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2023] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10968–10980, 2023.

[Liu *et al.*, 2024] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[Luo *et al.*, 2022] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022.

[Ma *et al.*, 2022] 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.

[Qi *et al.*, 2017] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature

learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30, 2017.

[Schuster *et al.*, 2015] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, Jan 2015.

[Wang *et al.*, 2022] 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.

[Wang *et al.*, 2023a] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10873–10883, 2023.

[Wang *et al.*, 2023b] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 3drp-net: 3d relative position-aware network for 3d visual grounding. *arXiv preprint arXiv:2307.13363*, 2023.

[Wu *et al.*, 2019] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[Wu *et al.*, 2023] 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.

[Yang *et al.*, 2020] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020.

[Yang *et al.*, 2021] 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.

[Yuan *et al.*, 2021] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021.

[Yuan *et al.*, 2022] 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.

[Yuan *et al.*, 2025a] Zhenlong Yuan, Cong Liu, Fei Shen, Zhaoxin Li, Jinguo Luo, Tianlu Mao, and Zhaoqi Wang. Msp-mvs: Multi-granularity segmentation prior guided multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9753–9762, 2025.

[Yuan *et al.*, 2025b] Zhenlong Yuan, Jinguo Luo, Fei Shen, Zhaoxin Li, Cong Liu, Tianlu Mao, and Zhaoqi Wang. Dvp-mvs: Synergize depth-edge and visibility prior for multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9743–9752, 2025.

[Zhang *et al.*, 2023] 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.

[Zhang *et al.*, 2024] 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*, volume 38, pages 7296–7304, 2024.

[Zhao *et al.*, 2021] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021.