Spatial Matters: Position-Guided 3D Referring Expression Segmentation

Yabing Wang[1] Zhuotao Tian[2] Le Wang[1*] Zheng Qin[1] Sanping Zhou[1]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Harbin Institute of Technology, Shenzhen

Abstract

3D Referring Expression segmentation (3D-RES) is an emerging field that segments 3D objects in point cloud scenes based on given referring expressions. Although existing methods have achieved substantial progress, they primarily focus on semantic cues and often overlook spatial relations, which are essential for segmenting the referred objects in complex 3D scenes, especially those containing multiple visually similar instances. In this paper, we propose Position3D, a novel approach that explicitly incorporates spatial relation modeling into 3D-RES. Specifically, we introduce a spatial-aware query generation module that constructs point proxies by aggregating local context and incorporating spatial relations, from which the most text-relevant are selected as queries. Furthermore, we design a position-guided deformable attention in the decoder, which progressively refines attention to concentrate on the target object under positional relationship guidance. Extensive experiments on two benchmark datasets, i.e.*, ScanRefer, and Multi3DRefer, validate the effectiveness of the proposed method Position3D* [1]

1. Introduction

3D Referring Expression Segmentation (3D-RES) aims to segment target objects within 3D point clouds based on natural language descriptions. By aligning linguistic cues with 3D visual-geometric features, 3D-RES enables intuitive, instruction-driven interaction with 3D environments, showing great potential for applications in AR/VR, embodied AI, and robotic manipulation.

Early approaches [15, 29, 47, 48] typically adopt a two-stage segmentation-then-matching paradigm, where a 3D instance segmentation network first generates object proposals, which are subsequently matched with the textual query to obtain the final mask. Although these methods achieve promising results, the two-stage paradigm heavily depends on the quality of the pre-segmentation and suffers from efficiency issues. To address this limitation, recent studies [5, 12, 45, 46] have shifted toward a one-stage paradigm, generally employing encoder–decoder architectures to perform multi-modal fusion and decode the referred object using object queries, similar to DETR [2], achieving state-of-the-art performance.

Unlike 2D images, 3D scenes are inherently more com-

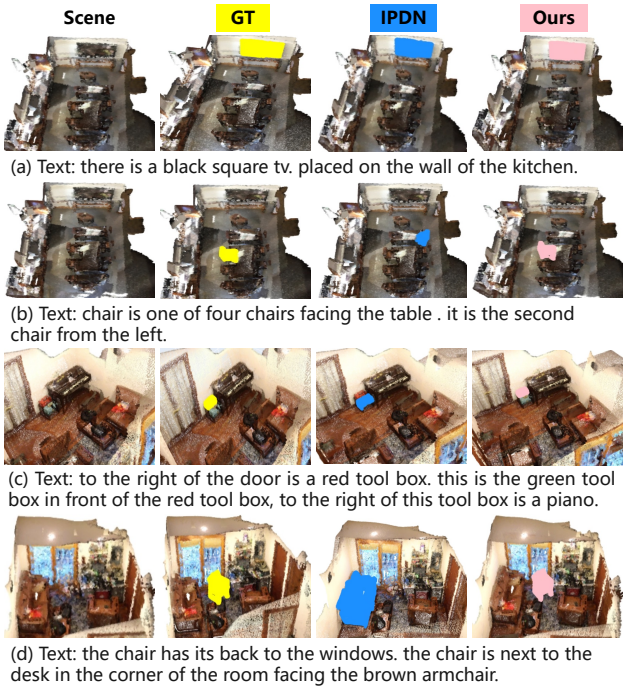


Figure 1. Illustration of challenges arising from the lack of spatial relation modeling. Existing methods (*e.g.* IPDN [5]) perform well when referring expressions are primarily driven by semantic cues in (a) but struggle with expressions that depend on spatial relationships in (b–d). Our method alleviates this limitation by explicitly modeling spatial relations.

*Corresponding author.
[1]https://github.com/LiJiaBei-7/Position3D

plex, containing a larger number of objects distributed throughout the space. As illustrated in Figure 1, referring expressions often describe objects not only by their appearance but also through their spatial relations with other objects (e.g., "the chair is next to the desk"). In particular, in scenes with multiple visually identical instances, such as the similar chairs shown in Figure 1 (b), spatial relational cues serve as a crucial signal for distinguishing the target objects. Therefore, modeling spatial relations effectively is key to achieving accurate 3D-RES.

Despite substantial progress in 3D-RES, existing methods primarily focus on semantic cues and largely neglect spatial awareness required for accurate segmentation of referred objects. While these methods perform well when objects can be distinguished by appearance or category, they often fail in scenarios requiring spatial reasoning. For instance, in Figure 1 (a), the referring expression relies mainly on semantic cues, which prior methods can handle effectively. In contrast, in Figures 1 (b–d), the expressions depend on spatial relations (e.g., relative positions), where these methods struggle due to the lack of explicit spatial modeling. This limitation highlights the need to explicitly model spatial relations in 3D-RES to enhance the model's spatial awareness.

To this end, we propose Position3D, a spatial-aware framework that explicitly captures spatial relations under position guidance. The network includes a spatial-aware query generation module that constructs point proxies by aggregating local context and embedding positional relations, thereby encoding both semantic and spatial cues. The most text-relevant proxies are then selected as decoder queries. In the decoder, a position-guided deformable attention mechanism is introduced to progressively refine attention toward the target region under positional guidance. This design enables the model to capture spatial relationships between queries and surrounding points, while focusing on the most informative regions in the 3D scene via sparse attention. Extensive qualitative and quantitative experiments on the ScanRefer[4] and Multi3DRefer[52] datasets validate the superior performance of Position3D.

Our main contributions are summarized as follows.

- We propose Position3D, a novel framework that explicitly models spatial relations to enhance spatial awareness and advance 3D referring expression segmentation.
- We design a spatial-aware query generation module to encode both semantic and spatial cues, enhancing spatial understanding. Additionally, we introduce position-guided deformable attention in the decoder, which progressively refines the attention to concentrate on the target object under the position guidance.
- We conduct extensive experiments on two 3D-RES benchmarks (ScanRefer and Multi3DRefer), achieving significant performance improvements and demonstrating the effectiveness of the proposed Position3D.

2. Related Work

2.1. 3D Visual Grounding

Driven by the success of multimodal learning [19, 23, 31–33, 37–43] in 3D scene understanding, 3D visual grounding (3DVG) aims to localize language-referenced objects within 3D point clouds. Existing methods are generally categorized into two-stage and one-stage frameworks. Two-stage methods [1, 11, 16, 49, 50] adopt a detect-then-match paradigm: pre-trained language models [7, 8, 27] encode the query, while pre-trained 3D detectors [24, 28] or segmenters [6, 18] generate object candidates; in the second stage, visual and textual features are aligned to localize the referred object. In contrast, one-stage methods [3, 10, 17, 25, 26, 34–36, 44] integrate object detection and feature extraction, allowing for direct identification of the target object. These methods typically rely on attention mechanisms to implicitly integrate multi-modal information. For instance, BUTD-DETR [17] utilizes transformer encoder–decoder layers to combine 3D visual representations with features from other modalities, while EDA [48] enhances fine-grained cross-modal alignment by decomposing the textual input.

2.2. 3D Referring Expression Segmentation

Compared with 3D Visual Grounding, 3D Referring Expression Segmentation (3D-RES) requires more fine-grained localization of the target object through mask prediction. TGNN [15] pioneers this task by adopting a two-stage framework that leverages graph neural networks to match candidate instances with textual descriptions. To further improve inference efficiency and segmentation accuracy, 3D-STMN [47] proposes an end-to-end one-stage framework with a superpoint-text matching mechanism driven by dependency-based reasoning. RefMask3D [12] extends this line of research by enabling more comprehensive multimodal feature interaction and understanding. MDIN [46] introduces the generalized 3D referring expression segmentation (3D-GRES) task, extending the capability to segment an arbitrary number of instances based on natural language instructions. Moreover, IPDN [5] incorporates multi-view 2D image cues into 3D scenes to compensate for spatial information loss. Despite these advances, most existing approaches fail to explicitly model spatial dependencies, which limits their ability to distinguish objects in complex or ambiguous scenes. To address this limitation, RG-SAN [45] decomposes textual expressions into object-centric representations, thereby facilitating more effective localization and spatial relation modeling. In this paper, we propose Position3D, a position-guided spatial relationship modeling framework that generates spatial-aware queries

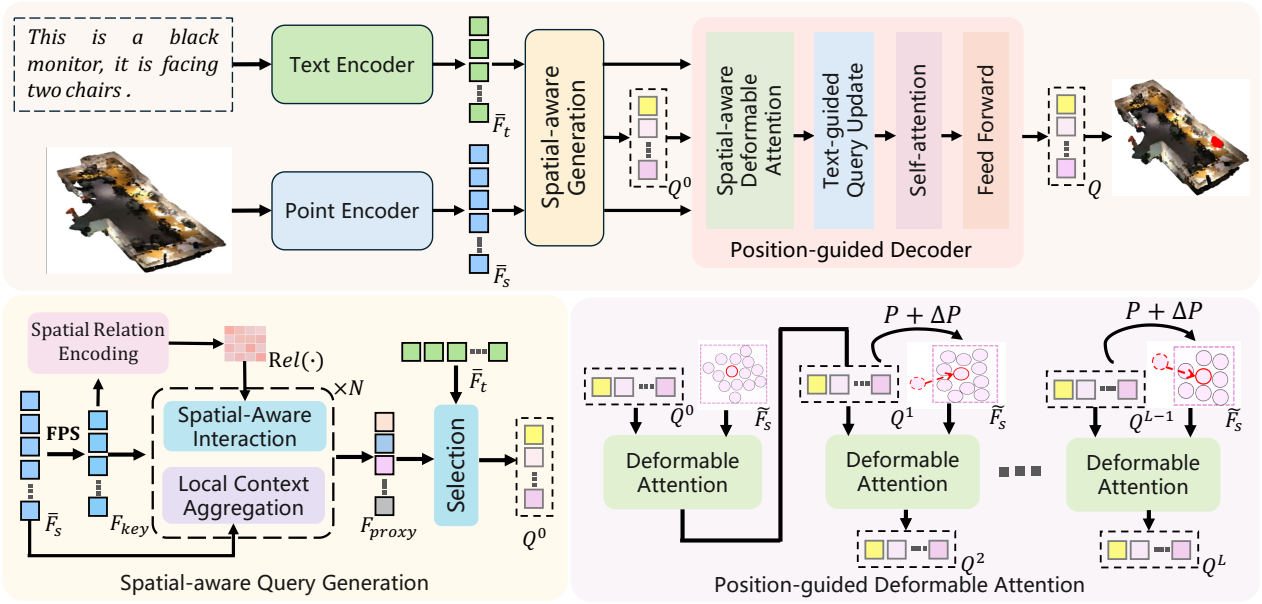


Figure 2. The overview of our proposed method, Position3D. To enhance spatial understanding, we introduce a spatial-aware query generation module that constructs point proxies by aggregating local contextual features and incorporating spatial relations to encode both semantic and positional cues. The most text-relevant proxies are then selected as decoder queries. In the decoder, a position-guided deformable attention is designed to iteratively refine attention toward the target object under explicit positional guidance. It operates in a sparse attention manner, enabling the model to focus on the most informative regions while effectively capturing spatial relations.

and guides the decoder to progressively localize the target object under positional relation modeling.

3. Method

Figure 2 shows an overview of our approach. We first describe the feature extraction in Section 3.1. Next, we present our spatial-aware query generation module in Section 3.2. In Section 3.3, we introduce the position-guided decoder. Finally, the training objectives are detailed in Section 3.4.

3.1. Feature Extraction

Given a textual description for the target object and a point cloud scene $P \in \mathbb{R}^{N_p \times 6}$ where N_p denotes the number of points and each point is represented by its 3D coordinates (x, y, z) and color (r, g, b), we extract both textual and visual feature representations.

Specifically, for textual features, we utilize a pre-trained RoBERTa [22] to extract word-level feature representations $F_t \in \mathbb{R}^{N_t \times D_t}$, where N_t and D_t denote the number of tokens and dimensionality of features, respectively. On the other hand, for visual features, we process the point cloud P using a sparse 3D U-Net [9] to extract point-wise feature maps $F_p \in \mathbb{R}^{N_p \times D_p}$. Following prior work [5, 46], we further generate N_s superpoints from the original point cloud and perform superpoint pooling on F_p to obtain the superpoint-level features $F_s \in \mathbb{R}^{N_s \times D_p}$.

Then, we employ two multi-layer perceptrons (MLPs) ϕ_t and ϕ_s to project the textual and visual feature a unified embedding space of dimension d:

$$\bar{\mathbf{F}}_t = \phi_t(F_t) \tag{1}$$

$$\bar{\mathbf{F}}_s = \phi_s(F_s) \tag{2}$$

Additionally, IPDN [5] indicates that multi-view 2D features can alleviate feature ambiguity caused by information loss or distortion in 3D representations. Therefore, the final visual features are obtained by the sum of the superpoint features and multi-view 2D features.

3.2. Spatial-aware Query Generation

We propose a spatial-aware query generation module that explicitly incorporates spatial cues into query construction, thereby providing stronger spatial awareness for the decoder. This module first constructs a set of spatial-aware point proxies by aggregating local context and introducing positional relationships into the interactions among proxies. These proxies jointly capture both semantic and geometric information in the scene. Subsequently, the most relevant ones are selected as queries based on textual features, serving as informative anchors for the decoder.

Spatial-aware Point Proxy. Given the superpoint features $\bar{\mathbf{F}}s$ and their 3D coordinates P_s, we first perform farthest point sampling (FPS) to select a subset of key points that are spatially representative and evenly dis-

tributed across the scene:

$$\mathbf{F}_{key} = \bar{\mathbf{F}}_s[\text{FPS}(P_s)] \tag{3}$$

Although FPS ensures broad geometric coverage, the sampled key points alone contain limited contextual and spatial information. To enrich their representation, we feed them into a stack of spatial context aggregation blocks, each consisting of two layers: a local context aggregation layer and a spatial-aware interaction layer.

Local Context Aggregation Layer. In order to model the local context, for each key point, we use KNN to identify its K_s nearest neighbors among the superpoints based on their 3D coordinates:

$$\mathbf{F}_n = \text{KNN}(\mathbf{F}_{key}, \bar{\mathbf{F}}_s; P_{key}, P_s; K_s) \tag{4}$$

where P_{key} denotes the 3D coordinates of the key points. Then, we derive the adaptive weights that selectively emphasize informative neighbors to capture local dependencies, resulting in context-enriched representations $\mathbf{F}_c$:

$$\mathcal{S} = \frac{\mathbf{F}_{key}\mathbf{F}_n^T}{\sqrt{D}} \tag{5}$$

$$\mathbf{F}_c = \text{Softmax}(\mathcal{S}) \cdot \mathbf{F}_n \tag{6}$$

Spatial-aware Interaction Layer. Beyond local context, we further encode global spatial relations among the key points. We compute their pairwise relative position matrix and map them into high-dimensional embeddings via an MLP $\phi_r(\cdot)$. This process is expressed as follows:

$$\text{Rel}(\mathbf{F}_{\text{key}})_{i,j} = \phi_r(x_i - x_j, y_i - y_j, z_i - z_j) \tag{7}$$

$$\text{with} \quad \forall i, j \in [1, N_{\text{key}}] \tag{8}$$

These spatial relation embeddings are then integrated into the self-attention mechanism to jointly model semantic and spatial dependencies:

$$\mathcal{A}_{\text{spa}} = \text{Rel}(\mathbf{F}_{\text{key}}) + \frac{\text{Q}(\mathbf{F}_{\text{key}})\,\text{K}(\mathbf{F}_{\text{key}})^\top}{\sqrt{d}} \tag{9}$$

$$\mathbf{F}_{\text{spa}} = \text{Softmax}(\mathcal{A}_{\text{spa}}\,\text{V}(\mathbf{F}_{\text{key}})) \tag{10}$$

Finally, the spatial-aware point proxies are obtained by fusing the local context features $\mathbf{F}_c$ and spatial-aware semantic features $\mathbf{F}_{\text{spa}}$ via an MLP $\phi_{\text{proxy}}(\cdot)$:

$$\mathbf{F}_{\text{proxy}} = \phi_{\text{proxy}}(\mathbf{F}_{\text{c}} + \mathbf{F}_{\text{spa}}) + \mathbf{F}_{\text{key}} \tag{11}$$

Query Selection. To generate spatially informed queries, we compute the similarity between the textual representations and the spatial-aware point proxies, and select the top-N_q most relevant proxies as decoder queries:

$$\mathcal{S}_{\text{proxy}} = \frac{1}{N_t}\sum_{i=1}^{N_t} \bar{\mathbf{F}}_{t,i} \cdot \mathbf{F}_{\text{proxy}}^\top \tag{12}$$

$$\mathbf{Q} = \text{TopK}(\mathcal{S}_{\text{proxy}}) \tag{13}$$

3.3. Position-Guided Decoder

Previous 3D-RES methods mainly emphasize semantic understanding of the target object, yet often overlook the intrinsic spatial relationships within the 3D scene. To address this limitation, we propose a position-guided deformable attention mechanism in the decoder, which progressively narrows the attention scope from global context to local regions via sparse attention, while explicitly modeling positional relationships between queries and point cloud features.

Position-guided Deformable Attention. To endow each query with explicit spatial awareness, we first predict a 3D center C^l for each query feature and iteratively refine it at each decoder layer:

$$\Delta P^l = \phi_{\text{offset}}(\mathbf{Q}^l), \tag{14}$$

$$C^l = C^{l-1} + \Delta P^l \tag{15}$$

where C^0 is initialized as the 3D coordinates of the selected proxy, and $\phi_{\text{offset}}(\cdot)$ is an MLP that maps the query feature into a 3D offset. This iterative refinement allows each query to progressively adjust its spatial focus, and its receptive field gradually narrows from global context to local regions.

After predicting C^l, we explicitly encode the geometric relationship between each query center and the super-points as a geometry prior:

$$\text{Rel}(C^l, P_s) = \phi_g(c_i^x - x_j, c_i^y - y_j, c_i^z - z_j) \tag{16}$$

$$\text{with} \quad i \in [1, N_q],\ \forall j \in [1, N_s] \tag{17}$$

where (x_j, y_j, z_j) denotes the coordinates of the j-th superpoint, and $\phi_g(\cdot)$ is an MLP that embeds the 3D relative positions into a learnable spatial embedding. This geometry prior provides each query with explicit spatial cues, guiding the subsequent attention computation.

While existing methods typically rely on global attention, in practice, only the target-related regions should receive concentrated attention, whereas most other points should be suppressed with negligible weights. To this end, we adopt a sparse attention mechanism, where for each query $\mathbf{Q}^i$, selects m^l nearest super-points based on the Euclidean distance to its predicted center:

$$\tilde{\mathbf{F}}_s^{(i,l)} = \left\{ f_j \,\middle|\, p_j \in \text{TopK}_{p_j \in P_s}\big(-\|\mathbf{c}_i^l - p_j\|_2, m^l\big) \right\} \tag{18}$$

where m^l denotes the number of selected super-points at layer l, which progressively decreases as l increases ($m^1 > m^2 > \cdots > m^L$) with $m^l \ll N_s$. As the query centers are iteratively refined, fewer points are selected in deeper layers, enabling the attention to gradually focus on more spatially precise neighborhoods. Attention is then computed

over these local points by jointly considering feature similarity and positional relations:

$$\mathcal{A}^l_{\text{pad}} = \text{Rel}(\mathbf{Q}^l, \tilde{\mathbf{F}}^l_s) + \frac{\text{Q}(\mathbf{Q}^l)\,\text{K}(\tilde{\mathbf{F}}^l_s)^\top}{\sqrt{d}} \tag{19}$$

$$\hat{\mathbf{Q}}^l = \text{Softmax}(\mathcal{A}^l_{\text{pad}}\,\text{V}(\tilde{\mathbf{F}}^l_s))) + \mathbf{Q}^l \tag{20}$$

Text-guided Query Update. Subsequently, the queries interact with textual features through cross-attention, allowing them to focus on 3D regions that are semantically aligned with the language description:

$$\mathbf{Q}^l_t = \text{Attention}(\hat{\mathbf{Q}}^l, \mathbf{F}_t, \mathbf{F}_t) \tag{21}$$

The text-enhanced queries $\mathbf{Q}^l_t$ are further refined via self-attention to capture contextual dependencies among queries, followed by a feed-forward network (FFN) to produce the updated query representation for the next layer:

$$\mathbf{Q}^s_t = \text{Attention}(\mathbf{Q}^l_t, \mathbf{Q}^l_t, \mathbf{Q}^l_t) \tag{22}$$

$$\mathbf{Q}^{l+1} = \text{FFN}(\mathbf{Q}^s_t) \tag{23}$$

3.4. Training Objective

Given the query obtained from the aforementioned, we feed it into the prediction head to generate mask $M \in \mathbb{R}^{N_q \times N_s}$ and confidence probability $Prob \in \mathbb{R}^{N_q \times 2}$, which indicate whether each query corresponds to the target instance:

$$M = \mathbf{Q} \cdot \phi_{\text{m}}(\bar{\mathbf{F}_s})^T \tag{24}$$

$$Prob = \phi_{\text{prob}}(\mathbf{Q}) \tag{25}$$

where ϕ_{m} and ϕ_{prob} are MLPs. Based on these predictions, our training objectives can be formulated as:

Mask Prediction. We adopt the Binary Cross-Entropy (BCE) loss and Dice loss as the training objective to supervise the model for precise mask prediction:

$$\mathcal{L}_{mask} = \text{BCE}(M, M_{\text{gt}}) + \text{DICE}(M, M_{\text{gt}}) \tag{26}$$

where M represents the predicted mask, and M_{gt} represents the ground truth mask.

Confidence Prediction. We employ the BCE loss to supervise the predicted confidence probability of each query:

$$\mathcal{L}_{cls} = \text{BCE}(Prob, Prob_{\text{gt}}) \tag{27}$$

where $Prob_{\text{gt}} \in 0, 1$ denotes whether each query corresponds to the target instance, with 1 indicating the presence of the target and 0 indicating its absence.

Center Prediction. To enable the query to more accurately localize the target center, we introduce an L1 loss to supervise the predicted center position of the query:

$$\mathcal{L}_{center} = \|C - C_{\text{gt}}\|_1, \tag{28}$$

where C_{gt} denotes the centroid of the target instance.

Cross-Modal Alignment. To bridge the gap between visual and textual modalities, we adopt the contrastive loss $\mathcal{L}_{contra}$ from EDA [48] to encourage alignment between text features and their matched query features.

Finally, the overall loss function is defined as follows:

$$\begin{aligned}\mathcal{L} = \lambda_{mask}\mathcal{L}_{mask} + \lambda_{cls}\mathcal{L}_{cls} \\ + \lambda_{center}\mathcal{L}_{center} + \lambda_{contra}\mathcal{L}_{contra}\end{aligned} \tag{29}$$

where λ denotes the hyperparameters that control the relative weight of each loss component.

4. Experiments

4.1. Datasets and Evaluation Metrics

ScanRefer. We evaluate our method using the ScanRefer dataset [4], which consists of 51,583 natural language expressions, encompassing 11,046 objects across 800 ScanNet scenes. The evaluation metrics include mean Intersection over Union (mIoU), Acc@0.25, and Acc@0.5.

Multi3DRefer. We evaluate our model on the Multi3DRefer dataset [52], which extends 3D-GRES by allowing each referring expression to correspond to any number of target objects, unlike traditional 3D-RES that assumes a single target. The dataset contains 61,926 expressions, including 51,583 adapted from ScanRefer, with 6,688 referring to zero objects, 13,178 to multiple objects, and the rest to a single object. Following the ScanRefer protocol, we report mIoU for all samples, assigning a score of 1 for zero-object expressions if the absence of a target is correctly predicted, and 0 otherwise.

4.2. Implementation Details

In our experiments, we adopt the PolyRL strategy with an initial learning rate of 0.0001 and a decay power of 4.0. The batch size is 16. We set the numbers of proxies and queries to 256 and 128, respectively, and use a 4-layer decoder. The hyperparameters m and K_s are set to $\{128, 64, 32, 16\}$ and 32, respectively. The loss weights λ_{mask}, λ_{cls}, $\lambda_{contrast}$, and λ_{center} are set to 1.0, 0.1, 0.1, and 0.5, respectively.

4.3. Comparisons with State-of-the-art Methods

As illustrated in Table 1, we compare our approach with other state-of-the-art methods on the ScanRefer benchmark. Our method demonstrates superior performance on both the "unique" and "multiple" subsets. In particular, the "multiple" subset contains ambiguous cases with several objects of the same category, where relying solely on semantic information is insufficient. By explicitly modeling spatial relations, our proposed Position3D achieves significant improvements in accurately segmenting the referred objects. These results highlight the effectiveness of incorporating

Table 1. Comparison with state-of-the-art methods on Unique, Multiple, and Overall settings.

Method	Venue	Unique (∼19%)			Multiple (∼81%)			Overall		
		0.25	0.5	mIoU	0.25	0.5	mIoU	0.25	0.5	mIoU
TGNN[15]	AAAI2021	69.3	57.8	50.7	31.2	26.6	23.6	38.6	32.7	28.8
InstanceRefer[51]	ICCV2021	81.6	72.2	60.4	29.4	23.5	21.5	40.2	33.5	30.6
3DRefTR [20]	Arxiv2023	89.6	77.0	–	52.3	43.7	–	57.9	48.7	41.2
X-RefSeg3D [29]	AAAI2024a	–	–	–	40.3	33.8	29.9	42.0	33.8	29.9
3D-STMN [47]	AAAI2024b	89.3	84.0	74.5	46.2	29.2	31.1	54.6	39.8	39.5
Reanson3D [14]	Arxiv2024	88.4	84.2	74.6	50.5	31.7	34.1	57.9	41.9	42.0
SegPoint [13]	ECCV2024	–	–	–	–	–	–	–	–	41.7
MCLN [30]	ECCV2024	89.6	78.2	–	**53.3**	45.9	–	58.7	50.7	44.7
RefMask3D [12]	ACMMM2024	89.6	84.7	–	48.1	40.1	–	55.9	49.2	44.2
MDIN [46]	ACMMM2024	91.0	87.2	76.7	50.1	44.9	41.4	56.3	51.1	48.3
IPDN [5]	AAAI2025	91.5	88.0	**78.6**	53.1	47.0	43.6	60.6	54.9	50.2
Position3D	CVPR2026	**92.0**	**88.9**	77.6	**54.1**	**48.7**	**44.7**	**61.5**	**56.1**	**51.0**

Table 2. The 3D-GRES results on Multi3DRefer. ZT, ST, and MT represent zero target, single target, and multiple targets respectively. The left and right sides of the "/" represent the situations with and without distractor objects, respectively.

Method	Acc@0.25				Acc@0.5				mIoU
	ZT	ST	MT	All	ZT	ST	MT	All	
ReLA [21]	36.2 / 72.7	48.3 / 83.4	73.0	61.8	36.2 / 72.7	20.4 / 65.5	42.4	37.4	42.8
M3DRef-CLIP [52]	39.2 / 81.6	50.8 / 77.5	66.8	55.7	39.2 / 81.6	29.4 / 67.4	41.0	37.5	37.4
3D-STMN [47]	42.6 / 76.2	49.0 / 77.8	68.8	60.4	42.6 / 76.2	24.6 / 69.2	43.9	40.9	43.0
MDIN [46]	**47.9** / 78.8	55.5 / 84.4	76.3	67.0	**47.9** / 78.8	29.5 / 71.7	46.8	44.7	47.5
IPDN [5]	39.4 / 84.1	61.5 / 88.9	79.6	71.5	39.4 / 84.1	34.7 / 79.5	52.1	50.0	51.7
Position3D	45.5 / **87.9**	**62.2 / 89.0**	**79.8**	**72.3**	45.5 / **87.9**	**37.8 / 82.3**	**53.4**	**52.7**	**53.2**

spatial relation, especially in challenging scenarios with multiple similar instances, and underscore the importance of spatial-aware modeling for 3D-RES.

We further evaluate our method on the Multi3DRefer benchmark, which involves more complex scenes with multiple target objects. As shown in Table 2, our model consistently outperforms existing methods across most settings, except for the zero-target scenario with distractors. In particular, in challenging scenarios with multiple objects and distractors, where accurately distinguishing the referred targets is difficult, our method achieves notable improvements by explicitly incorporating spatial cues. These results further validate that explicitly modeling spatial relations is crucial for improving 3D-RES performance in complex real-world environments.

4.4. Ablation Studies

In this section, we conduct ablation studies on ScanRefer to investigate the effectiveness of our proposed method.

Table 3. Ablation study of each component. "LCA" and "SAI" represent the local context aggregation layer and spatial-aware interaction layer described in Section 3.2, respectively. "SADA" denotes the position-guided deformable attention in Section 3.3.

LCA	SAI	SADA	Acc@0.25	Acc@0.5	mIoU
		✓	58.7	52.8	48.2
	✓	✓	58.6	53.3	48.5
✓		✓	59.2	53.8	48.9
✓	✓		59.5	54.4	49.4
✓	✓	✓	**61.5**	**56.1**	**51.0**

Effectiveness of the each component. To validate the effectiveness of each component in our method, we conduct an ablation study as shown in Table 3. In the first line, the queries are generated directly from the FPS-sampled key points without constructing point proxies leads to a sharp performance drop, highlighting the necessity of proxy-based query generation. Furthermore, removing either the

Table 4. Ablation study on the effectiveness of spatial relation modeling (i.e., $Rel(\cdot)$ function) in different components. "SR" denotes the spatial relation modeling.

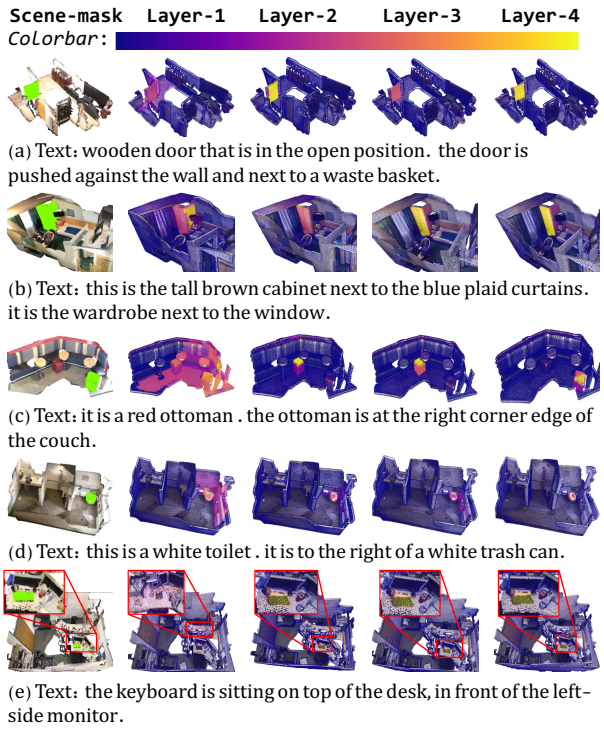
Method	Acc@0.25	Acc@0.5	mIoU
w/o SR in Query	59.4	54.2	49.2
w/o SR in Decoder	59.2	54.3	49.0
Full Model	**61.5**	**56.1**	**51.0**

Table 5. Ablation study on the number of super-points in the decoder's position-guided deformable attention. "all" indicates that all super-points features are used.

Method	Acc@0.25	Acc@0.5	mIoU
$\{all, all, all, all\}$	59.7	54.4	49.4
$\{128, 128, 128, 128\}$	59.9	54.5	49.6
$\{16, 16, 16, 16\}$	59.1	52.7	48.2
$\{all, 128, 64, 32\}$	60.1	55.0	50.0
$\{all, 64, 32, 16\}$	60.5	55.3	50.3
$\{128, 64, 32, 16\}$	**61.5**	**56.1**	**51.0**
$\{128, 64, 16, 8\}$	60.4	54.3	50.3

Table 6. Ablation study on the number of nearest neighbors in the local context aggregation layer.

k_s	Acc@0.25	Acc@0.5	mIoU
8	59.5	54.1	49.0
16	60.1	55.3	50.3
32	**61.5**	**56.1**	**51.0**
64	59.7	54.4	49.5



(a) Text: wooden door that is in the open position. the door is pushed against the wall and next to a waste basket.

(b) Text: this is the tall brown cabinet next to the blue plaid curtains. it is the wardrobe next to the window.

(c) Text: it is a red ottoman . the ottoman is at the right corner edge of the couch.

(d) Text: this is a white toilet . it is to the right of a white trash can.

(e) Text: the keyboard is sitting on top of the desk, in front of the left-side monitor.

Figure 3. Visualization of the position-guided deformable attention. Each example shows the progressive refinement of attention maps across four decoder layers (Layer 1–4) given the referring expression. The attention gradually focuses from the global context to the target region, demonstrating the effectiveness of our proposed position-guided deformable attention.

LCA or SAI layer further degrades performance, demonstrating that both local context aggregation and spatial-aware interaction are crucial for effective proxy construction. Finally, we relace the position-guided deformable attention with standard cross-attention results in a clear decline, validating the benefit of positional guidance in refining attention toward the target region.

Effectiveness of the spatial relation. As shown in Table 4, removing spatial relation modeling from either the query generation or decoder stage leads to a consistent performance drop across all metrics. Specifically, excluding spatial relation from the query generation module weakens the model's spatial awareness, resulting in a 2.1% drop in Acc@0.25 and 1.8% decrease in mIoU. Similarly, removing spatial relations from the decoder also degrades performance, highlighting the importance of spatial reasoning during decoding. In short, these results demonstrate that incorporating spatial relations at both stages effectively enhances the model's spatial understanding in 3D-RES.

Impact of the $\{m\}^L$ in the position-guided deformable attention. As shown in Table 5, we investigate the impact of varying the number of super-points, *i.e.* $\{m\}^L$, used in the decoder's position-guided deformable attention. Using all super-points leads to suboptimal performance due to redundant information, which results in sparse attention. Conversely, using too few super-points (e.g., $\{16, 16, 16, 16\}$) limits spatial coverage, resulting in degraded accuracy. The best performance is achieved with a progressively decreasing number of super-points $\{128, 64, 32, 16\}$, which effectively balances spatial coverage and feature compactness, enabling the model to focus more precisely on target regions.

Impact of the k_s in the local context aggregation layer. As shown in Table 6, we evaluate the impact of the number of nearest neighbors k_s used in the local context aggregation layer. When k_s is too small (e.g., 8), the model fails to capture sufficient local geometry, leading to performance degradation. Conversely, an excessively large k_s (*e.g.* 64) introduces redundant or noisy spatial information, slightly reducing accuracy. The best performance is achieved at $k_s = 32$, which provides a balanced trade-off between local feature richness and noise suppression. In our experiment, we set the parameter k_s to 32.

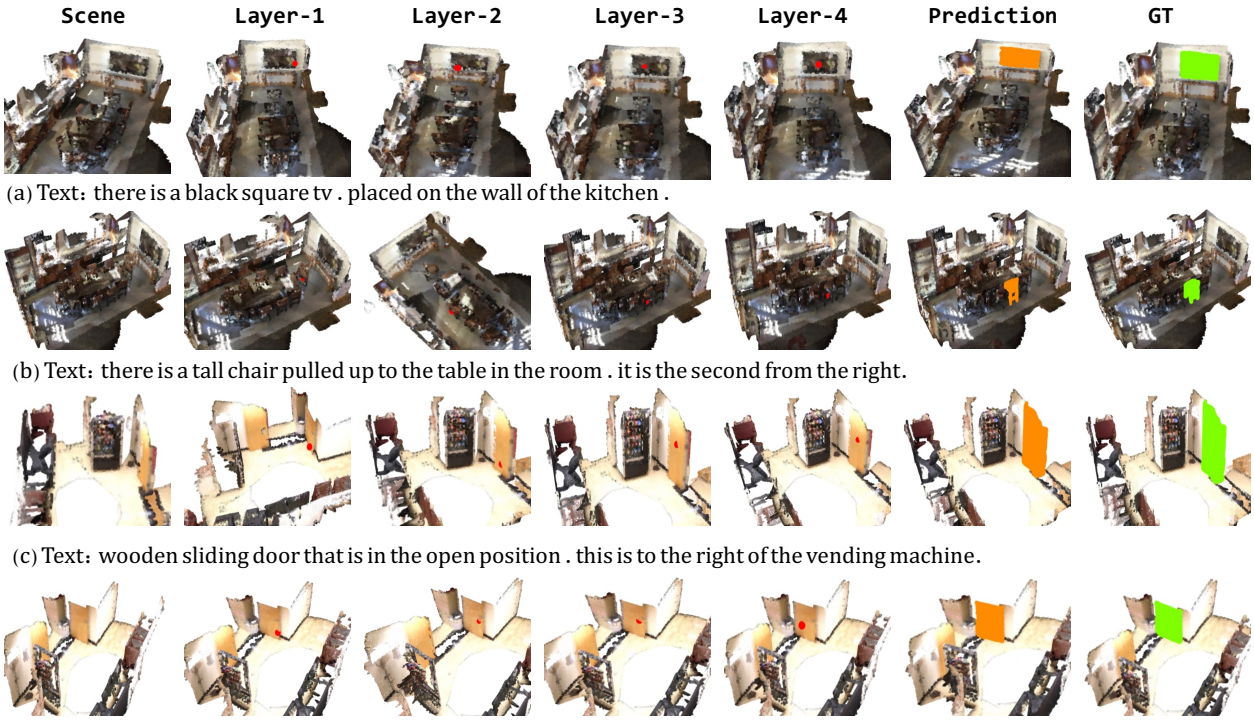


Figure 4. Visualization of the predicted centers (Layer 1–4) and final segmentation masks produced by our method, along with the ground-truth mask. The predicted centers are iteratively refined and gradually move toward the true center of the target object, demonstrating how the decoder progressively identifies the referred object and achieves accurate segmentation under positional guidance.

4.5. Qualitative Results

Visualization of the deformable attention. Figure 3 visualizes the position-guided deformable attention over four decoder layers. In the early layers (Layer 1–2), the attention primarily captures target-related context. As the decoder goes deeper, the attention gradually refines and becomes more focused on the target regions. For example, in case (e), the attention initially covers the "toilet" along with surrounding objects, such as the wall and trash can. Through iterative refinement, the attention precisely highlights the target object region. Moreover, even in cluttered 3D scenes with multiple similar objects, such as the two identical red ottomans" in case (c) or the crowded environment with a small target object in case (e), our method still performs well. It demonstrates that the position-guided deformable attention effectively leverages positional cues to progressively concentrate on relevant areas.

Visualization of the predicted center and segmentation mask. In Figure 4, we visualize the predicted centers from Layer 1 to Layer 4 in the decoder, along with the final segmentation mask. As the decoder goes deeper, the predicted centers gradually converge toward the true center of the referred object, illustrating how our position-guided decoder iteratively refines query localization. Moreover, even in scenes containing multiple visually similar objects, our method successfully segments the target. For example, in case (b), it accurately identifies the referred chair among several similar chairs based on positional relations. In cases (c) and (d), which belong to the same 3D scene, our method correctly segments two different doors according to their respective positions described in the referring expressions. These visualizations demonstrate that our approach can accurately identify and segment the target object through progressive refinement, even in scenes with multiple ambiguous objects.

5. Conclusion

In conclusion, we propose Position3D, a novel method that explicitly models spatial relations in 3D-RES. It first introduces a spatial-aware query generation module, which constructs point proxies to capture the semantic and spatial information. It then employs a position-guided deformable attention in the decoder, iteratively refining attention to focus on the target object while capturing spatial relations between the query with surrounding points. Experimental results demonstrate that Position3D achieves superior performance and effectively mitigates missegmentation in challenging scenarios with similar instances.

6. Acknowledgments

This work was supported in part by the National Key Research and Development Project under Grant 2024YFB4708100, the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM504, National Natural Science Foundation of China under Grants 62572384, U24A20325, 6212530562503384, and Key Research and Development Plan of Shaanxi Province under Grant 2024PT-ZCK-80. Thanks to Huawei Ascend servers for providing computing power support.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer, 2020. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[3] Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2024. 2

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

[5] Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, Danni Yang, and Xiaoshuai Sun. Ipdn: Image-enhanced prompt decoding network for 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2132–2140, 2025. 1, 2, 3, 6

[6] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021. 2

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2

[9] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 3

[10] Wenxuan Guo, Xiuwei Xu, Ziwei Wang, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Text-guided sparse voxel pruning for efficient 3d visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3666–3675, 2025. 2

[11] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15372–15383, 2023. 2

[12] Shuting He and Henghui Ding. Refmask3d: Language-guided transformer for 3d referring segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8316–8325, 2024. 1, 2, 6

[13] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pages 349–367. Springer, 2024. 6

[14] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In *International Conference on 3D Vision 2025*, 2025. 6

[15] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1610–1618, 2021. 1, 2, 6

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

[17] 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. 2

[18] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 2

[19] Jinlong Li, Cristiano Saltori, Fabio Poiesi, and Nicu Sebe. Cross-modal and uncertainty-aware agglomeration for open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19390–19400, 2025. 2

[20] 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. 6

[21] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 6

[22] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can

be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022. 3

[23] Xuexun Liu, Xu Xiaoxu, Jinlong Li, Qiudan Zhang, Xu Wang, Nicu Sebe, Ma Lin, et al. Less: Label-efficient and single-stage referring 3d instance segmentation. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. NeurIPS, 2024. 2

[24] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2949–2958, 2021. 2

[25] 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. 2

[26] Qihang Peng, Henry Zheng, and Gao Huang. Proxytransformation: Preshaping point cloud manifold with proxy attention for 3d visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24582–24592, 2025. 2

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[28] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2

[29] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4551–4559, 2024. 1, 6

[30] Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Multi-branch collaborative learning network for 3d visual grounding. In *European Conference on Computer Vision*, pages 381–398. Springer, 2024. 6

[31] Zheng Qin, Yabing Wang, Minghui Yang, Sanping Zhou, Ming Yang, and Le Wang. Embracing aleatoric uncertainty: Generating diverse 3d human motion. *arXiv preprint arXiv:2508.20604*, 2025. 2

[32] Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Yi Yuan, Jingdong Chen, and Le Wang. Humansense: From multimodal perception to empathetic context-aware responses through reasoning mllms. *arXiv preprint arXiv:2508.10576*, 2025.

[33] Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Zixin Zhu, Sanping Zhou, Ming Yang, and Le Wang. Versatile multimodal controls for expressive talking human animation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8253–8262, 2025. 2

[34] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 2

[35] Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Aware visual grounding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14056–14065, 2024.

[36] Austin Wang, ZeMing Gong, and Angel X Chang. Vigil3d: A linguistically diverse dataset for 3d visual grounding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30453–30475, 2025. 2

[37] Haoyu Wang, Le Wang, Sanping Zhou, Jingyi Tian, Zheng Qin, Yabing Wang, Gang Hua, and Wei Tang. Towards precise embodied dialogue localization via causality guided diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13350–13360, 2025. 2

[38] Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 422–433, 2022.

[39] Yabing Wang, Fan Wang, Jianfeng Dong, and Hao Luo. Cl2cm: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5651–5659, 2024.

[40] Yabing Wang, Shuhui Wang, Hao Luo, Jianfeng Dong, Fan Wang, Meng Han, Xun Wang, and Meng Wang. Dual-view curricular optimal transport for cross-lingual cross-modal retrieval. *IEEE Transactions on Image Processing*, 33:1522–1533, 2024.

[41] Yabing Wang, Zhuotao Tian, Qingpei Guo, Zheng Qin, Sanping Zhou, Ming Yang, and Le Wang. From mapping to composing: A two-stage framework for zero-shot composed image retrieval. *arXiv preprint arXiv:2504.17990*, 2025.

[42] Yabing Wang, Zhuotao Tian, Qingpei Guo, Zheng Qin, Sanping Zhou, Ming Yang, and Le Wang. Referencing where to focus: Improving visual grounding with referential query. *Advances in Neural Information Processing Systems*, 37: 47378–47399, 2025.

[43] Yabing Wang, Zhuotao Tian, Zheng Qin, Sanping Zhou, and Le Wang. Refdetector: A simple yet effective matching-based method for referring expression comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8033–8041, 2025. 2

[44] 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. 2

[45] Changli Wu, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, Rongrong Ji, et al. Rg-san: Rule-guided spatial awareness network for end-to-end 3d referring expression segmentation. *Advances in Neural Information Processing Systems*, 37:110972–110999, 2024. 1, 2

[46] Changli Wu, Yihang Liu, Jiayi Ji, Yiwei Ma, Haowei Wang, Gen Luo, Henghui Ding, Xiaoshuai Sun, and Rongrong Ji. 3d-gres: Generalized 3d referring expression segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7852–7861, 2024. 1, 2, 3, 6

[47] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5940–5948, 2024. 1, 2, 6

[48] 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. 1, 2, 5

[49] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4644–4653, 2019. 2

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

[51] 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. 6

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