PC-CrossDiff: Point-Cluster Dual-Level Cross-Modal Differential Attention for Unified 3D Referring and Segmentation

Wenbin Tan[1,2], Jiawen Lin[1], Fangyong Wang[3], Yuan Xie[4], Yong Xie[5]*, Yachao Zhang[1]*, Yanyun Qu[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, Xiamen, China
[2]School of Big Data, Tongren University, Tongren, China
[3]Hanjiang National Laboratory, Wuhan, China
[4]School of Computer Science and Technology, East China Normal University, Shanghai, China
[5]Department of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
wbtan@stu.xum.edu.cn, yongxie@njupt.edu.cn, yachaozhang@xmu.edu.cn

Abstract

3D Visual Grounding (3DVG) aims to localize the referent of natural language referring expressions through two core tasks: Referring Expression Comprehension (3DREC) and Segmentation (3DRES). While existing methods achieve high accuracy in simple, single-object scenes, they suffer from severe performance degradation in complex, multi-object scenes that are common in real-world settings, hindering practical deployment. Existing methods face two key challenges in complex, multi-object scenes: inadequate parsing of implicit localization cues critical for disambiguating visually similar objects, and ineffective suppression of dynamic spatial interference from co-occurring objects, resulting in degraded grounding accuracy. To address these challenges, we propose PC-CrossDiff, a unified dual-task framework with a dual-level cross-modal differential attention architecture for 3DREC and 3DRES. Specifically, the framework introduces: (i) Point-Level Differential Attention (PLDA) modules that apply bidirectional differential attention between text and point clouds, adaptively extracting implicit localization cues via learnable weights to improve discriminative representation; (ii) Cluster-Level Differential Attention (CLDA) modules that establish a hierarchical attention mechanism to adaptively enhance localization-relevant spatial relationships while suppressing ambiguous or irrelevant spatial relations through a localization-aware differential attention block. To address the scale disparity and conflicting gradients in joint 3DREC–3DRES training, we propose $\mathcal{L}_{\text{DGTL}}$, a unified loss function that explicitly reduces multi-task crosstalk and enables effective parameter sharing across tasks. Our method achieves state-of-the-art performance on the ScanRefer, NR3D, and SR3D benchmarks. Notably, on the Implicit subsets of ScanRefer, it improves the Overall@0.50 score by **+10.16%** for the 3DREC task, highlighting its strong ability to parse implicit spatial cues.

Code — https://github.com/tanwb/PC-CrossDiff.

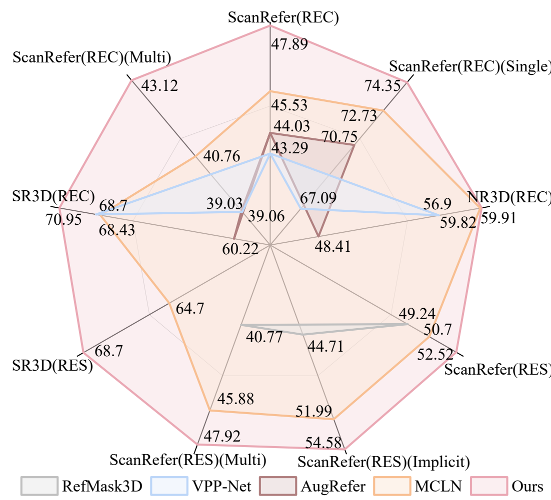

Figure 1: PC-CrossDiff vs. SOTA on 3DREC and 3DRES. ScanRefer (IoU@0.50), NR3D/SR3D (IoU@0.25).

Introduction

3D Visual Grounding (3DVG) enables machines to interpret natural language references in 3D scenes, with two core subtasks: 3D Referring Expression Comprehension (3DREC) for instance-level localization and 3D Referring Expression Segmentation (3DRES) for point-wise segmentation (Zhang et al. 2024; Xu et al. 2024; Wu et al. 2024). These tasks are complementary, as 3DREC identifies target objects while 3DRES delineates their boundaries. For example, warehouse robots must detect dynamic obstacles via 3DREC and segment their boundaries via 3DRES. However, addressing these tasks independently fails to leverage shared cross-modal information to disambiguate multi-object scenes, necessitating a unified framework that balances computational efficiency with robust grounding accuracy, especially in complex multi-object scenes.

This need is further emphasized by the limitations of existing single-task approaches. As shown in Fig. 1 on Scan-

*Corresponding Author.

Refer (Chen, Chang, and Nießner 2020), state-of-the-art methods achieve high accuracy only in rare single-object settings but degrade significantly in multi-object scenes due to visual ambiguity. Specifically, these methods achieve Unique@0.50>70% in single-object scenes (15% of cases), yet their performance drops below 46% in multi-object scenes (85% of cases), revealing limited robustness in complex multi-object scenes.

This performance gap reveals two fundamental limitations. (i) Inadequate parsing of implicit localization cues (such as "with magazines" or "beside the laptop"), which are crucial for disambiguating visually similar objects but remain underutilized due to the limited capacity of current frameworks. (ii) Insufficient suppression of dynamic spatial interference. In multi-object scenes, co-occurring objects introduce ambiguous spatial relations that existing methods aggregate indiscriminately, which amplifies noise and degrades performance.

Although extensive work has been conducted on single-task 3DREC and 3DRES, such as text–point cloud alignment (Wu et al. 2023a), attribute integration (Xu et al. 2024), and spatial modeling (Wang, Li, and Wang 2024), methods based on textual semantic parsing still primarily target explicit spatial relations (e.g., "A–top–B") and inadequately capture implicit cues that are critical for disambiguation in complex scenes. While graph-based models (Feng et al. 2021; Huang et al. 2021) and relational CNNs (Wang et al. 2019) capture spatial dependencies, they lack mechanisms to dynamically suppress interference from irrelevant objects. Although MCLN (Qian et al. 2024b) advances dual-task co-learning, the development of a unified framework robust to multi-object interference remains a significant challenge.

To bridge this gap, we propose PC-CrossDiff: a Point-Cluster Cross-modal Differential framework for unified 3D referring and segmentation. The proposed framework addresses the core limitations through two key innovations:

(1) For implicit cue parsing, we design a lightweight Point-Level Differential Attention (PLDA) module. Unlike conventional methods requiring explicit entity/relation parsing and graph construction, PLDA processes full sentences via cross-modal differential attention between point clouds and text to extract visually grounded features, including implicit cues, through a learnable weighting mechanism that adaptively focuses on localization-relevant features while suppressing irrelevant ones, enabling target disambiguation in multi-object scenes.

(2) For dynamic spatial filtering, we introduce Cluster-Level Differential Attention (CLDA). After modeling spatial relations among candidate points to capture global positional context and target-specific features, CLDA employs a Localization-Aware Differential Attention (LDA) block to selectively enhance localization-relevant spatial features while suppressing noise, enabling dynamic spatial relation filtering for robustness.

Furthermore, we propose the Dual-Geometry Task-Harmonized Loss ($\mathcal{L}_{\text{DGTL}}$) to resolve scale disparity and gradient conflicts in joint 3DREC–3DRES training, reducing multi-task crosstalk and enabling effective parameter sharing across tasks.

In summary, our contributions are:

- We propose PC-CrossDiff, a unified cross-modal differential learning framework operating at both point and cluster levels. By synergistically integrating local perception with global spatial filtering, it enables competitive performance on both 3DREC and 3DRES within a single architecture.
- We design PLDA to extract implicit spatial cues directly from full-text descriptions via cross-modal differential attention, supporting accurate disambiguation in complex multi-object scenes; CLDA at the cluster level enables dynamic spatial relation filtering by selectively enhancing localization-relevant features through a localization-aware attention mechanism, improving robustness in complex multi-object scenes.
- Extensive experiments on ScanRefer, NR3D/SR3D, and their Implicit/Multiple subsets demonstrate that PC-CrossDiff achieves state-of-the-art performance on both 3DREC and 3DRES, significantly outperforming prior approaches.

Related work

3D Visual Grounding

Advances in 3DVG are driven by the growing demand for indoor applications. Current methods can be categorized into two subtasks: 3DREC (Zhang et al. 2024; Xu et al. 2024) and 3DRES (Wu et al. 2024; Qian et al. 2024a). 3DREC is typically implemented via two-stage architectures, which use 3D object detectors or RPNs for proposal generation before cross-modal alignment (Zhang et al. 2024; Achlioptas et al. 2020), or via single-stage approaches that directly establish visual-textual correspondences (Luo et al. 2022; Jain et al. 2022). Research on 3DRES remains relatively limited.

Recent progress focuses on enhancing textual semantic understanding in 3D scenes (Feng et al. 2021; Wu et al. 2024b) and on improving cross-modal feature correspondences (Wu et al. 2023a; Qian et al. 2024a; Wu et al. 2023b, 2025). Transformer-based frameworks (Jain et al. 2022; He et al. 2021; Zhao et al. 2021; Wu et al. 2023a) excel through attention-driven visual-textual fusion, alongside approaches in point cloud segmentation (Zhang et al. 2025, 2021). Crucially, unified modeling of both tasks remains rare; MCLN (Qian et al. 2024b) is the only method that attempts joint modeling of 3DREC and 3DRES, yet it suffers severe performance degradation in multi-object scenes, limiting its practical deployment. This limitation motivates our design of a robust, unified framework.

Spatial Positional Relationship Extraction

In recent years, modeling spatial relationships in point clouds has become a focal research area for context-aware feature learning (Wu et al. 2024a). Early CNN-based approaches (Wang et al. 2019; Liu et al. 2019a) encode positional relations through local geometric feature extraction. Subsequent graph-based methods (Wang, Lin, and Wu 2024;

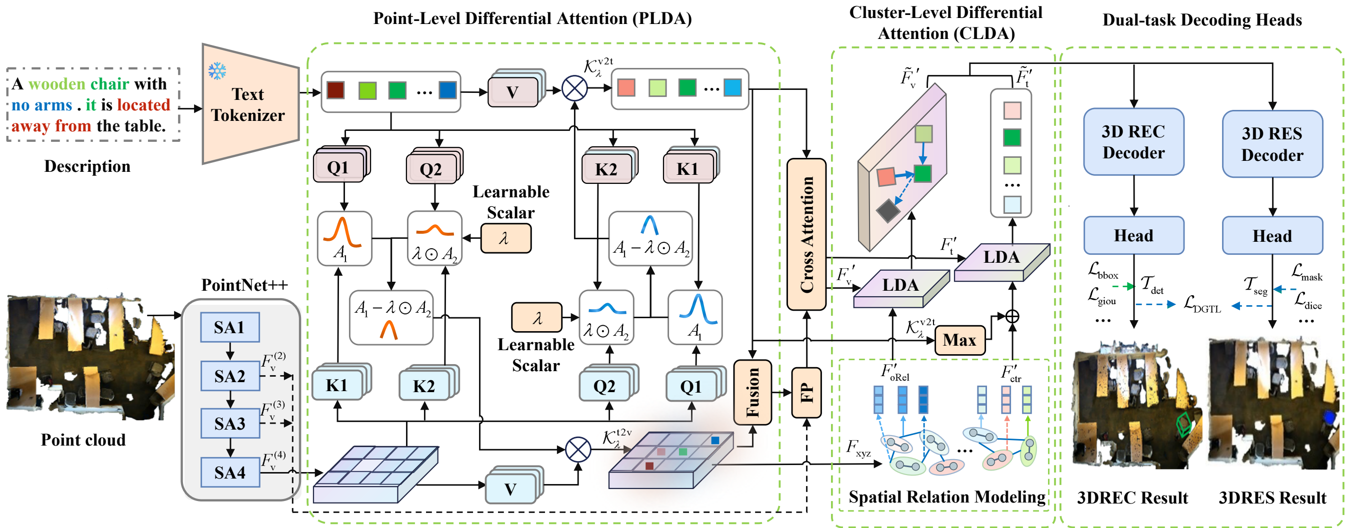

Figure 2: The framework of PC-CrossDiff. Text and visual features are input into the PLDA module to extract implicit localization cues and enhance point-level visual features. Simultaneously, the coordinates F_{xyz} are fed into the CLDA module for Spatial Relation Modeling (SRM), modeling cluster-level relations. These features are enhanced by LDA and integrated into the aligned features F_{v}' and F_{t}', yielding $\tilde{F}_{\mathrm{v}}'$ and $\tilde{F}_{\mathrm{t}}'$. The enhanced features are decoded by 3DREC and 3DRES heads to generate predictions, with $\mathcal{L}_{\mathrm{DGTL}}$ facilitating task synergy.

Feng et al. 2021) capture inter-object spatial dependencies. Recent advances further enhance spatial understanding; for example, TGNN (Huang et al. 2021) improves local contextual modeling with graph neural networks, while MA2TransVG (Xu et al. 2024) and G^{3}LQ (Wang, Li, and Wang 2024) leverage multi-attribute interactions and geometric perception, respectively.

However, existing methods lack dynamic filtering of spatial interference, making it difficult to distinguish hierarchical relationships in multi-object scenes. We address this limitation by introducing a cluster-level differential attention mechanism.

Methods

Overview of PC-CrossDiff

To enhance localization robustness in multi-object scenes, PC-CrossDiff integrates PLDA and CLDA into a unified differential learning framework. This two-level design synergistically combines local perception with global spatial filtering, enabling precise visual-text alignment and dynamic suppression of interference from co-occurring objects. As shown in Fig. 2, PC-CrossDiff adopts a DETR-like architecture with dual decoders for joint 3DREC and 3DRES.

The input point cloud $\mathcal{P}$ and text T are processed via PointNet++ (Qi et al. 2017a) and pretrained RoBERTa (Liu et al. 2019b) feature extractors. This yields textual features $F_{\mathrm{t}} \in \mathbb{R}^{l_{\mathrm{t}} \times d_{\mathrm{emb}}}$ (l_{t}: tokens, d_{emb}: embedding dim) and high-level visual features $F_{\mathrm{v}}^{(4)} \in \mathbb{R}^{n_{\mathrm{sem}} \times d_{\mathrm{sem}}}$ (n_{sem}: points, d_{sem}: semantic dim). The PLDA module computes fine-grained cross-modal localization features $F_{\mathrm{PLDA}}' \in \mathbb{R}^{n_{\mathrm{sem}} \times d_{\mathrm{sem}}}$ incorporating implicit cues. These features are fused with intermediate PointNet++ features $F_{\mathrm{v}}^{(3)}$ and $F_{\mathrm{v}}^{(2)}$ via upsampling and aligned cross-modally to produce F_{v}' and F_{t}'. To mitigate interference from multi-level spatial relationships in F_{v}' within multi-object scenes, F_{t}', F_{v}', and positional coordinates $F_{\mathrm{xyz}} \in \mathbb{R}^{n_{\mathrm{P}} \times 3}$ (n_{p}: points) are processed by the CLDA module to enhance localization-related features. The refined features feed into dual decoders.

Point-Level Differential Attention

Inspired by the Differential Transformer (Ye et al. 2024), we propose the Point-Level Differential Attention (PLDA) module to extract implicit spatial cues from descriptions, which enhances discriminative features via a single-layer bidirectional cross-modal differential attention mechanism. The module comprises two symmetric components: (1) *Text-to-Visual*: Enhances point-level visual features using textual localization references; (2) *Visual-to-Text*: Extracts implicit localization cues from text guided by visual features.

Visual-to-Text Cross-modal Differential Attention. This layer projects high-level visual features $F_{\mathrm{v}}^{(4)}$ to queries (Q) and textual features T to keys (K)/values (V) via learnable matrices $W^Q, W^K, W^V \in \mathbb{R}^{d_{\mathrm{sem}} \times d_{\mathrm{sem}}}$:

$$Q = F_{\mathrm{v}}^{(4)} W^Q, \quad K = T W^K, \quad V = T W^V. \tag{1}$$

The projections are reshaped into multi-head structures (head dimension $d_{\mathrm{h}} = d_{\mathrm{sem}}/(2N_{\mathrm{h}})$, $N_{\mathrm{h}} = 8$):

$$\begin{aligned} Q &= \mathrm{reshape}\left(Q, [n_{\mathrm{sem}}, 2N_{\mathrm{h}}, d_{\mathrm{h}}]\right)^{\intercal} = [Q_1, Q_2], \\ K &= \mathrm{reshape}\left(K, [l_{\mathrm{t}}, 2N_{\mathrm{h}}, d_{\mathrm{h}}]\right)^{\intercal} = [K_1, K_2], \\ V &= \mathrm{reshape}\left(V, [l_{\mathrm{t}}, N_{\mathrm{h}}, 2d_{\mathrm{h}}]\right)^{\intercal}. \end{aligned} \tag{2}$$

Then dual attention kernels are computed as:

$$A_1 = \mathrm{softmax}\left(\frac{Q_1 K_1^{\intercal}}{\sqrt{d_{\mathrm{h}}}}\right), A_2 = \mathrm{softmax}\left(\frac{Q_2 K_2^{\intercal}}{\sqrt{d_{\mathrm{h}}}}\right). \tag{3}$$

A learnable parameter λ suppresses irrelevant attention:

$$\lambda = \exp\left(\sum(\lambda_{\mathbf{q}_1} \odot \lambda_{\mathbf{k}_1})\right) - \exp\left(\sum(\lambda_{\mathbf{q}_2} \odot \lambda_{\mathbf{k}_2})\right),$$

where $\lambda_{\mathbf{q}_1}, \lambda_{\mathbf{k}_1}, \lambda_{\mathbf{q}_2}, \lambda_{\mathbf{k}_2} \in \mathbb{R}^{d_\mathrm{h}}$ are learnable vectors. The visual-to-text cross-modal differential attention is then computed as:

$$\mathrm{DiffAttn}(F_\mathrm{v}^{(4)}, T) = (A_1 - \lambda \odot A_2)V. \tag{4}$$

After layer normalization and linear projection, we obtain the visual-to-text cross-modal semantic feature $\mathcal{K}_{\mathrm{v2t}}$, encoding implicit localization cues for subsequent fusion.

Text-to-Visual Cross-modal Differential Attention. This component enhances point-level visual features by using textual features T as queries (Q) and $F_\mathrm{v}^{(4)}$ as keys (K)/values (V). The differential attention operation (Eqs. 1–4) is applied with the query and key roles interchanged. The learnable λ suppresses irrelevant visual features guided by the textual features, yielding enhanced point-level visual features $\mathcal{K}_{\mathrm{t2v}}$. For more details, see the Appendix.

Cluster-Level Differential Attention

To address the challenge of interference from irrelevant objects in multi-object scenes, we propose the Cluster-Level Differential Attention (CLDA) module for dynamic spatial filtering. CLDA first models cluster-level spatial relations with the Spatial Relation Modeling (SRM) block and subsequently applies two Localization-Aware Differential Attention (LDA) blocks to selectively enhance target-relevant features while suppressing noise, thereby achieving robust, dynamic spatial filtering in the presence of multi-object interference (Fig. 3).

Spatial Relation Modeling (SRM). Given position coordinates $F_{\mathrm{xyz}} \in \mathbb{R}^{n_\mathrm{p} \times 3}$ from F'_v (where F_v and F'_v share identical coordinates), SRM applies farthest point sampling (FPS) (Qi et al. 2017b) to select N_{clust} candidate centroids. KNN clustering partitions points into N_{clust} clusters of size $S_{\mathrm{clust}} = n_\mathrm{p}/N_{\mathrm{clust}}$ ($n_\mathrm{p} = 1024$).

An encoder processes intra-cluster points to obtain relative spatial relationships $F_{\mathrm{iRel}} \in \mathbb{R}^{N_{\mathrm{clust}} \times d_{\mathrm{mod}}}$. Inter-cluster spatial relationships are extracted via DGCNN (Wang et al. 2019):

$$F'_{\mathrm{oRel}} = \mathrm{DGCNN}(F_{\mathrm{iRel}}, \mathbf{O}) \in \mathbb{R}^{N_{\mathrm{clust}} \times d_{\mathrm{mod}}}, \tag{5}$$

where $\mathbf{O} \in \mathbb{R}^{N_{\mathrm{clust}} \times 3}$ represents cluster centroids, and candidate target region features are computed using a linear projection $F'_{\mathrm{ctr}} = \mathbf{O}\mathbf{W}_\mathrm{o} + \mathbf{b}_\mathrm{o}$, where $\mathbf{W}_\mathrm{o} \in \mathbb{R}^{3 \times d_{\mathrm{mod}}}$.

Localization-Aware Differential Attention (LDA). LDA implements a two-stage process: (1) feature filtering via single-layer unidirectional differential attention, and (2) feature enhancement via multi-head self-attention (Yu et al. 2022), as shown in Fig. 3.

For spatial relationship filtering, aligned visual features F'_v serve as queries (Q), while inter-cluster relationships F'_{oRel} act as keys (K) and values (V). The computation follows Eqs. 1–4:

$$\mathrm{DiffAttn}(F'_\mathrm{v}, F'_{\mathrm{oRel}}) = (A_1 - \lambda \odot A_2)V. \tag{6}$$

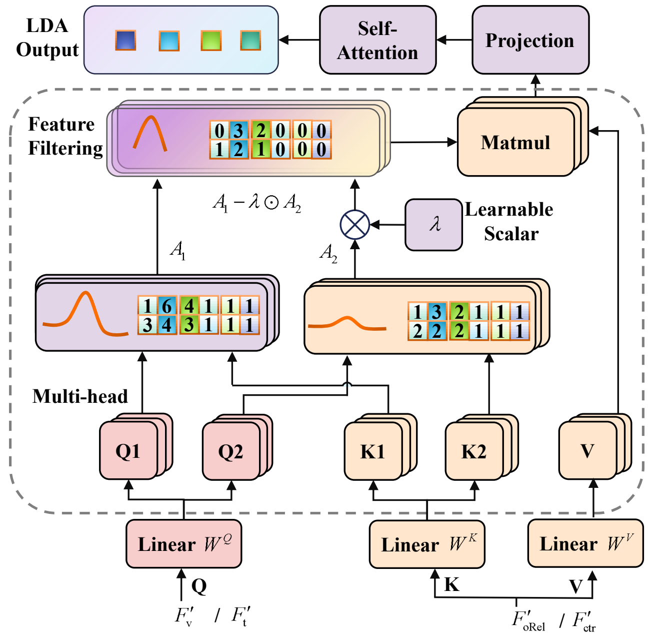

Figure 3: LDA block. With F'_v as query, DiffAtten (Eq. 6) extracts relevant relations from F'_{oRel} while suppressing irrelevant ones. Features then pass self-attention to yield $\tilde{F}'_\mathrm{v}$.

This extracts and enhances localization-relevant features from F'_{oRel} into F'_v, followed by self-attention refinement, yielding spatially enhanced features $\tilde{F}'_\mathrm{v}$. Similarly, using F'_t as queries (Q) and candidate regions F'_{ctr} as keys (K) and values (V), the computation is performed via Eqs. 1–4. The computation selects localization-relevant target region features from F'_{ctr} and enhances F'_t with them, refining target region features via self-attention. See Appendix for details.

Training Loss

We propose the Dual-Geometry Task-Harmonized Loss ($\mathcal{L}_{\mathrm{DGTL}}$) for joint 3DREC and 3DRES optimization:

Dual-Geometry Consistency. We enforce bidirectional alignment between localization ($\mathbf{B}_{\mathrm{loc}}$) and segmentation ($\mathbf{M}$) geometries to establish structural coherence:

$$\mathcal{L}_{\mathrm{geom}} = \zeta(t)\left[\mathcal{L}_{\mathrm{IoU}}(\mathbf{B}_{\mathrm{loc}}, \mathbf{B}_{\mathrm{mask}}) + \mathcal{L}_{\mathrm{Dice}}(\mathbf{M}, \mathbf{M}_{\mathrm{box}})\right],$$

where $\zeta(t)$ implements a linear warm-up, $\mathcal{L}_{\mathrm{IoU}}$ and $\mathcal{L}_{\mathrm{Dice}}$ are IoU and Dice losses, and $\mathbf{B}_{\mathrm{mask}}$ (mask-derived box) with $\mathbf{M}_{\mathrm{box}}$ (box-derived mask) form a self-supervised consistency loop via internal predictions.

Contribution Balancing. We balance contributions via $w_i = \max(e^{-v_i}, \lambda_i)$, where v_i is learnable and λ_i is an empirical threshold from MCLN (Qian et al. 2024b), ensuring meaningful task ratios.

Cross-task Harmonization. To harmonize optimization directions between detection ($\mathcal{T}_{\mathrm{det}}$) and segmentation ($\mathcal{T}_{\mathrm{seg}}$), we penalize conflicting gradients as:

$$\mathcal{P} = \eta(t) \cdot \sum_{i \in \mathcal{T}_{\mathrm{det}},\, j \in \mathcal{T}_{\mathrm{seg}}} \rho_{ij} \cdot \mathbb{I}\left[\cos(\theta_{ij}) < \tau\right] \cdot (\tau - \cos(\theta_{ij})),$$

where $\cos(\theta_{ij})$ measures gradient alignment, $\eta(t)$ is a linear decay factor, τ represents a threshold, $\mathbb{I}[\cdot]$ is the indicator function, and ρ_{ij} denotes task correlations.

The full loss $\mathcal{L}_{\text{DGTL}}$ (denoted as $\mathcal{L}$) integrates all components:

$$\mathcal{L} = \underbrace{\sum_{i \in \mathcal{T}_{\text{det}}} w_i \mathcal{L}_i + \sum_{j \in \mathcal{T}_{\text{seg}}} w_j \mathcal{L}_j + \mathcal{P}}_{\text{Task-Harmonized}} + \underbrace{\mathcal{L}_{\text{geom}}}_{\text{Dual-Geometry}}, \quad (7)$$

with dynamically learned weights w_i and $\mathcal{L}_{\text{geom}}$ enforcing spatial consistency.

Experiments

Experimental Setup

We evaluate PC-CrossDiff against the state-of-the-art multi-task baseline MCLN (Qian et al. 2024b) on ScanRefer (Chen, Chang, and Nießner 2020), NR3D/SR3D (Achlioptas et al. 2020), and their challenging subsets (Implicit and Multiple), under identical protocols for fair comparison across 3DREC and 3DRES tasks. Segmentation annotations strictly follow ScanNet's official standards (Dai et al. 2017). All experiments run on NVIDIA 3090 GPUs. We employ AdamW optimizer with learning rates of 2×10^{-3} for visual encoders and 2×10^{-4} for other layers. The temperature coefficient $\tau = 0.5$ and the number of network layers $L = 6$. Evaluation metrics and remaining hyperparameters align with MCLN's standards.

Quantitative Comparisons

Performance on ScanRefer (3DRES). We conduct comparative experiments on the ScanRefer dataset with TGNN(Huang et al. 2021) and current state-of-the-art methods: X-RefSeg3D(Qian et al. 2024a), 3D-STMN(Wu et al. 2024), SegPoint(He et al. 2025), MCLN(Qian et al. 2024b), and RefMask3D(He and Ding 2024). As shown in Table 1, our PC-CrossDiff achieves superior performance, achieving 60.41%, 52.52%, and 46.39% on Overall@0.25, Overall@0.50, and mIoU metrics, respectively. Specifically, PC-CrossDiff achieves improvements of 20.08%, 5.81%, and 4.54% over X-RefSeg3D, 3D-STMN, and RefMask3D on Overall@0.25. Similarly, it improves by 18.75%, 12.72%, and 3.28% on Overall@0.50. These results demonstrate that PC-CrossDiff achieves significant performance advantages in 3DRES tasks while maintaining a high inference speed of 4.07 FPS. This is attributed to the lightweight PLDA and CLDA modules, which enhance the model's robustness in multi-object environments via improved text understanding (especially implicit localization cues) and spatial filtering.

Performance on Multiple and Implicit Subsets. To evaluate PC-CrossDiff's performance in multi-object scenes, we conduct experiments on the Multiple subset of ScanRefer. To further assess its capacity to capture implicit localization cues, we construct an Implicit subset comprising 501 referring expressions with implicit references from the ScanRefer validation set (see Appendix). We use the model weights trained under the settings of Table 1 for inference, and the results are summarized in Table 2. On the Multiple@0.50 metric, PC-CrossDiff achieves improvements of 24.31%, 18.72%, and 7.15% over X-RefSeg3D (Qian et al. 2024a), 3D-STMN (Wu et al. 2024), and RefMask3D (He and Ding 2024), respectively.

On the Implicit subset, our method obtains a 17.25% gain over X-RefSeg3D and a 14.26% improvement over 3D-STMN in Implicit@0.50. Notably, it outperforms MCLN (baseline) by +10.16% in 3DREC. These results demonstrate PC-CrossDiff's superior ability to handle complex multi-object scenes and effectively exploit implicit spatial cues.

Method	**Venue**	**Overall**		**mIoU**	**FPS**
		0.25	**0.50**		
TGNN	AAAI'21	37.50	31.40	27.80	2.85
X-RefSeg3D	AAAI'24	40.33	33.77	29.94	2.03
3D-STMN	AAAI'24	54.60	39.80	39.50	4.14
SegPoint	ECCV'24	-	-	41.70	-
MCLN	ECCV'24	58.70	50.70	44.72	4.81
RefMask3D	ACMMM'24	55.87	49.24	44.86	1.65
Ours	/	**60.41**	**52.52**	**46.39**	4.07

Table 1: Comparison results of 3DRES on the ScanRefer dataset (inference: batchsize=1, on an NVIDIA 3090 GPU).

Method	**Multiple**		**Implicit**	
	0.25	**0.50**	**0.25**	**0.50**
3DRES				
X-RefSeg3D	28.16	23.61	43.71	37.33
3D-STMN	46.20	29.20	57.07	40.32
RefMask3D	48.09	40.77	49.30	44.71
MCLN	53.28	45.88	59.16	51.99
Ours	**55.33**	**47.92**	**62.15**	**54.58**
3DREC				
MCLN	51.96	40.76	54.18	38.45
Ours	**53.59**	**43.12**	**60.76**	**48.61**

Table 2: Joint Evaluation of 3DRES and 3DREC across Multiple and Implicit Subsets in the ScanRefer Dataset.

Inference Speed. PC-CrossDiff was evaluated under identical hardware and input settings. As shown in Table 1, our method achieves 4.07 FPS, outperforming X-RefSeg3D (Qian et al. 2024a) and RefMask3D (He and Ding 2024) in terms of both localization accuracy and inference speed. This efficiency stems from the lightweight design of the PLDA and CLDA modules; for instance, PLDA uses a single layer of cross-modal differential attention with only 2.10M parameters, introducing a negligible increase in model size and complexity. See the Appendix for details.

Performance on ScanRefer (3DREC). Table 3 compares PC-CrossDiff with recent single-stage and two-stage methods. In the two-stage setting, we compare with MVT (Huang et al. 2022), 3D-SPS (Luo et al. 2022), BUTD-DETR (Jain et al. 2022), ViL3DRel (Chen et al. 2022), 3D-VLP (Jin

Method	Venue	Multiple		Overall	
		0.25	**0.50**	**0.25**	**0.50**
Two-stage Model					
MVT	CVPR'22	31.92	25.26	40.80	33.26
3D-SPS	CVPR'22	40.32	29.82	48.82	36.98
BUTD-DETR	ECCV'22	44.73	33.97	50.42	38.60
ViL3DRel	NIPS'22	40.30	30.71	47.94	37.73
3D-VLP	CVPR'23	43.51	33.41	51.41	39.46
EDA	CVPR'23	49.13	37.64	54.59	42.26
3DRefTR-SP	ICCV'23	50.07	38.65	55.45	43.48
3D-VisTA	ICCV'23	43.70	39.10	50.60	45.80
VPP-Net	CVPR'24	50.53	39.03	55.65	43.29
G3-LQ	CVPR'24	51.48	40.80	56.90	45.58
Dual-Task					
MCLN	ECCV'24	51.96	40.76	57.17	45.53
Ours	/	**53.59**	**43.12**	**58.47**	**47.89**
Single-stage Model					
3D-SPS	CVPR'22	39.48	29.61	47.65	36.43
BUTD-DETR	ECCV'22	44.20	32.81	49.76	37.05
EDA	CVPR'23	48.11	36.82	53.83	41.70
G3-LQ	CVPR'24	50.23	39.72	55.95	44.72
AugRefer	AAAI'25	49.96	39.06	55.68	44.03
TSP3D	CVPR'25	-	-	56.45	46.71
TSP3D*	CVPR'25	50.66	41.30	55.82	45.69
Dual-task					
MCLN	ECCV'24	49.72	38.41	54.30	42.64
Ours	/	**51.59**	**42.02**	**57.37**	**46.68**

Table 3: Comparison results of single-stage and two-stage 3DREC on the ScanRefer dataset (*: Reproduction).

et al. 2023), EDA (Wu et al. 2023a), 3D-VisTA (Zhu et al. 2023), VPP-Net (Shi, Wu, and Lee 2024), G3-LQ (Wang, Li, and Wang 2024), MA2TransVG (Xu et al. 2024), and MCLN. PC-CrossDiff achieves 58.47% and 47.89% on Overall@0.25 and Overall@0.50, respectively, surpassing all baselines. On Multiple@0.50, it improves by 2.36% over MCLN. In the single-stage setting, we compare with AugRefer (Wang et al. 2025), TSP3D (Guo et al. 2025), and other state-of-the-art models. PC-CrossDiff exceeds MCLN and AugRefer by 3.61% and 2.96% on Multiple@0.50, and improves by 4.04% and 2.65% on Overall@0.50. Notably, its Overall@0.50 score surpasses several two-stage models, including VPP-Net, G3-LQ, and MCLN.

Performance on SR3D/NR3D. Following MCLN's evaluation protocol, we compare PC-CrossDiff with 3D-SPS (Luo et al. 2022), MVT (Huang et al. 2022), BUTD-DETR (Jain et al. 2022), LAR (Bakr, Alsaedy, and Elhoseiny 2022), EDA (Wu et al. 2023a), VPP-Net (Shi, Wu, and Lee 2024), and MCLN. As shown in Table 4, PC-CrossDiff achieves state-of-the-art performance on both benchmarks. Specifically, on SR3D's 3DREC task, PC-CrossDiff outperforms VPP-Net and MCLN by 2.25% and 2.52%, respectively, with a 2.71% improvement over VPP-Net on NR3D,

Method	Venue	SR3D	NR3D
3D-SPS	CVPR'22	62.60	51.50
MVT	CVPR'22	64.50	55.10
BUTD-DETR	ECCV'22	65.60	49.10
LAR	NIPS'22	59.60	48.90
EDA	CVPR'23	68.10	52.10
VPP-Net	CVPR'24	68.70	56.90
AugRefer	AAAI'25	60.22	48.41
Dual-Task			
MCLN	ECCV'24	68.43	59.82
Ours	/	**70.95**	**59.91**

Table 4: 3DREC performance on SR3D and NR3D (Acc@0.25IoU).

demonstrating strong generalization capability.

Component			3DREC		3DRES		
$\mathcal{L}_{\mathrm{DGTL}}$	**PLDA**	**CLDA**	**0.25**	**0.50**	**0.25**	**0.50**	**mIoU**
w/o	w/o	w/o	56.76	44.86	58.43	49.66	43.85
✓	w/o	w/o	56.95	45.50	58.49	50.86	44.78
✓	✓	w/o	57.56	46.20	59.19	50.99	45.12
✓	w/o	✓	57.69	46.18	59.61	51.88	45.68
✓	✓	✓	**58.47**	**47.89**	**60.41**	**52.52**	**46.39**

Table 5: Ablation study of $\mathcal{L}_{\mathrm{DGTL}}$, PLDA, and CLDA on 3DREC/3DRES (✓: enabled; w/o: disabled).

Ablation Studies

All ablation studies are conducted on the ScanRefer dataset to evaluate performance on both 3DREC and 3DRES tasks.

To systematically evaluate the $\mathcal{L}_{\mathrm{DGTL}}$, PLDA, and CLDA modules, we conduct ablation studies on the ScanRefer benchmark. As shown in Table 5, introducing $\mathcal{L}_{\mathrm{DGTL}}$ alone improves the Overall@0.50 score by 1.2% on the 3DRES task, validating the necessity of the loss in mitigating learning imbalance. Adding PLDA alone boosts performance by 1.33%, validating its ability to extract implicit localization cues through bidirectional attention and refine point-level features. Similarly, incorporating CLDA alone yields gains of 2.22%, demonstrating its effectiveness in dynamic spatial relation modeling and noise suppression for robust localization in multi-object scenes. When all three components are combined, the model achieves optimal performance with notable improvements on the Multiple and Implicit subsets (Appendix 3.1), demonstrating synergy across components.

To evaluate the robustness of localization accuracy to cluster size, we analyze its impact. As shown in Table 6, when the cluster size increases from $S_{\mathrm{clust}} = 8$ to $S_{\mathrm{clust}} = 16$, performance varies by less than 0.32% for 3DREC@0.25. This stability indicates that our functional decoupling design effectively mitigates clustering-induced errors, with additional details provided in the Appendix.

We evaluate feature transfer from the PLDA module to the

CLDA		3DREC		3DRES		
N_{clust}	S_{clust}	**0.25**	**0.50**	**0.25**	**0.50**	**mIoU**
256	4	56.11	46.37	57.83	50.88	44.66
128	8	58.15	47.09	60.21	52.40	46.16
64	16	**58.47**	**47.89**	**60.41**	**52.52**	**46.39**
32	32	57.36	47.01	59.25	51.53	45.54

Table 6: Ablation study of cluster configurations (N_{clust} and S_{clust}) on 3DREC and 3DRES tasks.

CLDA module's Max block (Fig. 2) and analyze differential attention (DiffAtten, Eq. 4) through two experiments: (i) Comparing original visual features $F_{\text{v}}^{(4)}$ with $\mathcal{K}_{\lambda}^{\text{t2v}}$ obtained via text-guided DiffAtten; (ii) Comparing original text features F_{t} with $\mathcal{K}_{\lambda}^{\text{v2t}}$ obtained via vision-guided DiffAtten.

Component	3DREC		3DRES		
	0.25	**0.50**	**0.25**	**0.50**	**mIoU**
$F_{\text{v}}^{(4)}$	56.94	45.98	58.87	50.97	44.94
$\mathcal{K}_{\lambda}^{\text{t2v}}$	58.06	46.92	60.25	51.84	46.13
F_{t}	57.93	46.71	60.02	52.42	46.27
$\mathcal{K}_{\lambda}^{\text{v2t}}$	**58.47**	**47.89**	**60.41**	**52.52**	**46.39**

Table 7: Comparison of localization accuracy when integrating different feature configurations from PLDA into CLDA.

As shown in Table 7, both $F_{\text{v}}^{(4)}$ and F_{t}, when processed by DiffAtten, significantly improve localization accuracy over their original counterparts, validating the effectiveness of cross-modal differential attention. Notably, text-based $\mathcal{K}_{\lambda}^{\text{v2t}}$ achieves optimal performance through two mechanisms: (i) F_{t} provides more precise target information; (ii) PLDA enhances discriminative localization cues in text while suppressing irrelevant features and improves feature discriminability.

Component	3DREC		3DRES		
	0.25	**0.50**	**0.25**	**0.50**	**mIoU**
$\star$	56.64	45.81	58.54	50.63	44.85
PLDA+CLDA	**58.47**	**47.89**	**60.41**	**52.52**	**46.39**

Table 8: Performance comparison of our dual-level differential attention and standard cross-attention ($\star$: standard Transformer-based (Vaswani et al. 2017)).

Table 8 shows that replacing the proposed dual-level differential attention with standard Transformer attention (Vaswani et al. 2017) degrades Overall@0.50 by 2.08% and 1.89% on 3DREC and 3DRES, respectively. This degradation reveals that standard attention fails to preserve implicit semantic cues and suppress irrelevant spatial interference. In contrast, our framework explicitly addresses these

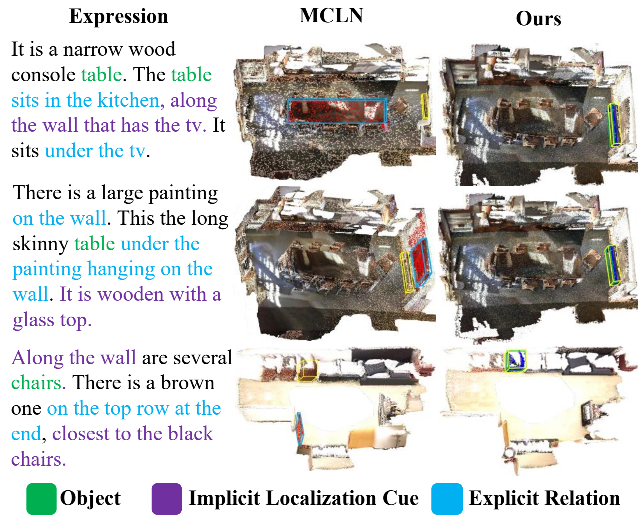

Figure 4: Comparison of PC-CrossDiff (ours) and MCLN (baseline) for 3DREC and 3DRES with implicit cues in complex multi-object scenes. GT boxes are yellow.

issues by leveraging PLDA for extracting implicit localization cues and CLDA for dynamic spatial filtering, validating the need for a dual-level design.

Qualitative Results

We compare PC-CrossDiff with the state-of-the-art dual-task method MCLN (Qian et al. 2024b) on complex multi-object scenes. As shown in Fig. 4, MCLN mislocalizes targets due to missing implicit cues (Rows 1–2). In contrast, PC-CrossDiff successfully captures target locations, particularly for implicit or occluded instances (Rows 1–3). The superior performance stems from two key mechanisms: bidirectional cross-modal interaction in PLDA mitigates semantic ambiguity in implicit localization, while dynamic spatial weighting in CLDA resolves spatial relationship ambiguity. These components jointly enable robust and accurate localization in complex multi-object scenes.

Conclusion

In this paper, we propose PC-CrossDiff, a novel dual-hierarchy optimization framework, to address the challenges of inadequate implicit semantic parsing and spatial interference in 3DVG for multi-object scenes. PLDA introduces a text-point cloud bidirectional differential attention mechanism to adaptively extract implicit localization cues, mitigating semantic ambiguity caused by explicit relation parsing in conventional approaches. CLDA enhances localization robustness through dynamic spatial relationship modeling with suppression of irrelevant interference, enabling efficient reasoning in multi-object scenes. A unified multi-task model integrating 3DREC and 3DRES achieves state-of-the-art performance on ScanRefer and NR3D/SR3D benchmarks, demonstrating its effectiveness in handling multi-object scenes.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176224, Grant 62176092, Grant 62222602, Grant 62306165, and Grant 62172234; in part by the Fundamental Research Funds for the Central Universities under Grant 20720250031; in part by the Science and Technology on Sonar Laboratory under Grant 2024-JCJQ-LB-32/07; in part by the Guizhou Provincial Department of Science and Technology under Grant Qian Ke He QN [2025] 337; and in part by the Tongren Municipal Science and Technology Bureau under Grant Tongshi [2025] 54.

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 422–440. Springer International Publishing. ISBN 978-3-030-58452-8.

Bakr, E.; Alsaedy, Y.; and Elhoseiny, M. 2022. Look around and Refer: 2d Synthetic Semantics Knowledge Distillation for 3d Visual Grounding. 35: 37146–37158.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, ECCV, 202–221. Springer International Publishing. ISBN 978-3-030-58565-5.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022. Language Conditioned Spatial Relation Reasoning for 3D Object Grounding. 35: 20522–20535.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Niessner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 5828–5839.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. Free-Form Description Guided 3D Visual Graph Network for Object Grounding in Point Cloud. In *CVPR*, CVPR, 3722–3731.

Guo, W.; Xu, X.; Wang, Z.; Feng, J.; Zhou, J.; and Lu, J. 2025. Text-Guided Sparse Voxel Pruning for Efficient 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3666–3675.

He, D.; Zhao, Y.; Luo, J.; Hui, T.; Huang, S.; Zhang, A.; and Liu, S. 2021. TransRefer3D: Entity-and-Relation Aware Transformer for Fine-Grained 3D Visual Grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 2344–2352. Association for Computing Machinery. ISBN 978-1-4503-8651-7.

He, S.; and Ding, H. 2024. RefMask3D: Language-Guided Transformer for 3D Referring Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 8316–8325. Association for Computing Machinery. ISBN 979-8-4007-0686-8.

He, S.; Ding, H.; Jiang, X.; and Wen, B. 2025. SegPoint: Segment Any Point Cloud via Large Language Model. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 349–367. Springer Nature Switzerland. ISBN 978-3-031-72670-5.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2021. Text-Guided Graph Neural Networks for Referring 3D Instance Segmentation. 35(2): 1610–1618.

Huang, S.; Chen, Y.; Jia, J.; and Wang, L. 2022. Multi-View Transformer for 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15524–15533.

Jain, A.; Gkanatsios, N.; Mediratta, I.; and Fragkiadaki, K. 2022. Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, ECCV, 417–433. Springer Nature Switzerland. ISBN 978-3-031-20059-5.

Jin, Z.; Hayat, M.; Yang, Y.; Guo, Y.; and Lei, Y. 2023. Context-Aware Alignment and Mutual Masking for 3d-Language Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10984–10994.

Liu, Y.; Fan, B.; Xiang, S.; and Pan, C. 2019a. Relation-Shape Convolutional Neural Network for Point Cloud Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, 8895–8904.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, J.; Fu, J.; Kong, X.; Gao, C.; Ren, H.; Shen, H.; Xia, H.; and Liu, S. 2022. 3D-SPS: Single-Stage 3D Visual Grounding via Referred Point Progressive Selection. In *CVPR*, CVPR, 16454–16463.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017a. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qian, Z.; Ma, Y.; Ji, J.; and Sun, X. 2024a. X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured Cross-Modal Graph Neural Networks. 38(5): 4551–4559.

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

Shi, X.; Wu, Z.; and Lee, S. 2024. Viewpoint-Aware Visual Grounding in 3D Scenes. In *CVPR*, CVPR, 14056–14065.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.;

Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, S.; Lin, Y.; and Wu, Y. 2024. Omni-Q: Omni-Directional Scene Understanding for Unsupervised Visual Grounding. In *CVPR*, CVPR, 14261–14270.

Wang, X.; Zhao, N.; Han, Z.; Guo, D.; and Yang, X. 2025. AugRefer: Advancing 3D Visual Grounding via Cross-Modal Augmentation and Spatial Relation-based Referring. 39(8): 8006–8014.

Wang, Y.; Li, Y.; and Wang, S. 2024. G^3-LQ: Marrying Hyperbolic Alignment with Explicit Semantic-Geometric Modeling for 3D Visual Grounding. In *CVPR*, CVPR, 13917–13926.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic Graph CNN for Learning on Point Clouds. 38(5): 146:1–146:12.

Wu, C.; Ma, Y.; Chen, Q.; Wang, H.; Luo, G.; Ji, J.; and Sun, X. 2024. 3D-STMN: Dependency-Driven Superpoint-Text Matching Network for End-to-End 3D Referring Expression Segmentation. 38(6): 5940–5948.

Wu, Y.; Cheng, X.; Zhang, R.; Cheng, Z.; and Zhang, J. 2023a. EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding. In *CVPR*, CVPR, 19231–19242.

Wu, Y.; Xing, M.; Zhang, Y.; Luo, X.; Xie, Y.; and Qu, Y. 2024a. UniDSeg: Unified Cross-Domain 3D Semantic Segmentation via Visual Foundation Models Prior. *Advances in Neural Information Processing Systems*, 37: 101223–101249.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; Fan, J.; Shi, Z.; and Qu, Y. 2023b. Cross-Modal Unsupervised Domain Adaptation for 3D Semantic Segmentation via Bidirectional Fusion-then-Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 490–498. Association for Computing Machinery. ISBN 979-8-4007-0108-5.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; Peng, K.; and Qu, Y. 2025. Fusion-then-Distillation: Toward Cross-modal Positive Distillation for Domain Adaptive 3D Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2024b. CLIP2UDA: Making Frozen CLIP Reward Unsupervised Domain Adaptation in 3D Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 8662–8671. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0686-8.

Xu, C.; Han, Y.; Xu, R.; Hui, L.; Xie, J.; and Yang, J. 2024. Multi-Attribute Interactions Matter for 3D Visual Grounding. In *CVPR*, CVPR, 17253–17262.

Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; and Wei, F. 2024. Differential Transformer. In *The Thirteenth International Conference on Learning Representations*.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-BERT: Pre-Training 3D Point Cloud Transformers With Masked Point Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19313–19322.

Zhang, Y.; Hu, R.; Li, R.; Qu, Y.; Xie, Y.; and Li, X. 2024. Cross-Modal Match for Language Conditioned 3D Object Grounding. 38(7): 7359–7367.

Zhang, Y.; Lan, Y.; Xie, Y.; Li, C.; and Qu, Y. 2025. Cross-cloud consistency for weakly supervised point cloud semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15520–15528.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3DVG-Transformer: Relation Modeling for Visual Grounding on Point Clouds. In *CVPR*, CVPR, 2928–2937.

Zhu, Z.; Ma, X.; Chen, Y.; Deng, Z.; Huang, S.; and Li, Q. 2023. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2911–2921.