LaSA-Net: A Language-Guided Network for Outdoor Generalized 3D Referring Expression Segmentation

Lingfei Ma[1] **Bin Liu**[1*] **Wen Li**[2] **Wentao Sun**[3]
Chaorui Liu[1] **Haiyan Guan**[4] **Jonathan Li**[1*]
[1]East China Normal University [2]University of Bristol
[3]University of Waterloo [4]Nanjing University of Information Science and Technology
https://github.com/NJSkate/LaSA_Net

Abstract

3D Referring Expression Segmentation (3D-RES) aims to segment target objects in point clouds via language. However, extending 3D-RES from indoor scenes to large-scale outdoor environments introduces three major challenges, i.e., large-scale geospatial complexity lacking in indoor benchmarks, sparse-target perception overwhelmed by massive geometric clutter and distractors, and context-induced query dilution, where highly relation-intensive expressions cause globally-initialized queries to be absorbed by salient reference objects. To address these issues, we introduce UrbanRefer, the first outdoor 3D-GRES benchmark with 193 urban road point cloud scenes and 7,840 textual descriptions for complex geospatial reasoning. We further propose LaSA-Net, a Language-guided Semantic Segmentation Network. First, to enhance sparse-target perception, LaSA-Net incorporates pixel-level dense features from DINOv3 as 2D priors and fuses them with 3D features through depth-consistent verification. Then, to alleviate query dilution, we design a semantic-adaptive query generation module that applies text-weighted farthest point sampling to focus on text-relevant regions, and performs multi-granularity sampling to maintain both local grounding and global semantic consistency. Finally, a reliability-aware decoder adaptively suppresses noisy language cues through cross-modal reliability-gated injection, generating precise prompts for accurate mask prediction. Extensive experiments demonstrate that LaSA-Net achieves 52.5%, 55.9%, and 47.9% mIoU on the ScanRefer, Multi3DRefer, and UrbanRefer datasets, surpassing prior state-of-the-art methods by 2.3%, 4.2%, and 2.8% respectively.

1 Introduction

3D Referring Expression Segmentation (3D-RES) aims to segment specific instances in point clouds according to natural language descriptions, serving as a fundamental capability for autonomous driving, robot navigation, and human–machine interaction [1]. Early 3D-RES methods adopt a two-stage paradigm, which gradually evolves into more efficient single-stage query-based architectures. However, traditional 3D-RES is built upon a single-target assumption [2], where each instruction is expected to correspond to a uniquely existing object. This assumption deviates from real-world scenarios involving ambiguity, multiplicity, and absence of targets. To address this limitation, recent works extend the task to Generalized 3D Referring Expression Segmentation (3D-GRES) [3], enabling models to handle multi-target and zero-target cases within a unified framework.

Despite these advances, existing 3D-RES and 3D-GRES methods are primarily developed and evaluated in indoor environments. In such settings, scenes are spatially compact, object distributions are relatively balanced, and spatial relations are dominated by local ego-centric cues such as "left

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of" or "next to" [4]. It remains unclear whether these methods can generalize to large-scale outdoor environments, which exhibit extended spatial ranges, dense clutter, and complex relational structures.

In this work, we go beyond the conventional indoor setting and formulate a new problem termed Outdoor 3D-GRES (3D-OGRES). In this setting, models are required to perform language-guided segmentation over spatially extensive urban scenes. As summarized in Tab. 1, this setting differs fundamentally from indoor scenarios in three aspects. First, the scene scale shifts from compact room-level environments to large-scale open urban spaces. Second, spatial relations evolve from local ego-centric descriptions to global geospatial reasoning over cardinal directions, topological layouts, and long-range dependencies. Third, target distributions become highly imbalanced, with sparse objects of interest embedded within dense same-category distractors. To support this setting, we introduce **UrbanRefer**, a large-scale outdoor 3D-GRES benchmark containing 193 outdoor point cloud scenes and 7,840 annotated descriptions, with an emphasis on geospatial relations and multi-object contextual reasoning.

Table 1: Comparison between indoor 3D-GRES and the Outdoor 3D-GRES setting.

Aspect	Indoor 3D-GRES	Outdoor 3D-GRES (Ours)
Scene scale	$30.7\,\mathrm{m}^2$	$547.7\,\mathrm{m}^2$
Spatial relation type	Local ego-centric relations	Global geospatial reasoning
Target distribution	0.81%	0.26%

The 3D-OGRES problem introduces two fundamental challenges [5]. (1) **Sparse-target perception.** Unlike indoor scenes with relatively balanced object distributions, outdoor point clouds are dominated by massive geometric clutter. Targets of interest, such as pedestrians and poles, are often small, sparse, and structurally incomplete, while severe distractor interference arises from visually similar objects. Relying solely on sparse 3D geometry provides insufficient fine-grained cues for reliable discrimination, making true targets easily overwhelmed by surrounding background structures. (2) **Context-induced query dilution**. Referring expressions in outdoor scenes are often concise but heavily rely on contextual spatial relations, while explicit appearance cues remain limited. Under such conditions, conventional globally initialized queries tend to be attracted to salient but irrelevant reference objects. Meanwhile, decoders that treat language cues uniformly cannot effectively suppress semantic noise, causing attention to gradually shift toward contextual regions and weakening the focus on true targets, which degrades precise mask prediction.

To address these challenges, we propose **LaSA-Net**, a language-guided segmentation framework tailored for 3D-OGRES. (1) To overcome sparse-target perception, we incorporate complementary dense visual priors from 2D foundation models to compensate for the lack of fine-grained cues in sparse 3D geometry. Specifically, LaSA-Net introduces pixel-level features from DINOv3 and achieves precise 2D–3D fusion through depth-consistent verification, enabling enriched target-aware representations. (2) To mitigate query dilution, we design a semantic-adaptive query generation module and a reliability-aware decoder. The query generation module applies text-weighted farthest point sampling (FPS) to guide query initialization toward text-relevant regions, followed by multi-granularity sampling to preserve both local grounding and global semantic consistency. The decoder dynamically evaluates the reliability of language priors and suppresses noisy cross-modal cues through reliability-gated injection, thereby generating precise and robust prompts for accurate mask prediction. In summary, our main contributions are as follows:

- We formulate Outdoor 3D-GRES as a new problem setting and introduce UrbanRefer, a large-scale outdoor benchmark for urban road scenes, characterized by geospatial reasoning, absolute directional relations, and highly imbalanced target distributions.
- We propose LaSA-Net, which addresses sparse-target perception via 2D-prior-enhanced feature fusion and mitigates context-induced query dilution through semantic-adaptive query generation and reliability-aware decoding.
- Extensive experiments demonstrate that LaSA-Net achieves state-of-the-art performance on both indoor and outdoor datasets. Especially on the outdoor scene, LaSA-Net achieves 47.9% mIOU, surpassing previous state-of-the-art methods by 2.8%.

2 Related Work

3D Referring Expression Comprehension. The 3D Referring Expression Comprehension (3D-REC) task aims to predict a bounding box for the object referred to by a given expression. Existing methods for 3D-REC can be broadly divided into two categories: two-stage methods [6, 7, 8], which first employ a detection model to generate candidate regions and then perform semantic matching between the text and these regions, For example, TransRefer3D [9] improved grounding by modeling entity attributes and inter-object relations, while InstanceRefer [10] enhanced grounding through multi-level contextual reasoning. Single-stage methods [11, 12] reduce the reliance on independent candidate generation and adopt an end-to-end joint optimization paradigm. For example, 3D-SPS [13] progressively selected referred points to focus localization on query-relevant geometry, while Proxy-Transformation [14] reshaped point cloud features with proxy attention to strengthen cross-modal representation learning. Recent studies have further extended 3D-REC to more practical settings, including embodied reference understanding [15] and outdoor autonomous driving scenes [16, 17]. Specifically, CityRefer [18] and CityAnchor [19] studies language-guided single-object localization in large-scale urban point clouds, using UAV photogrammetry and colored city-scale point clouds, respectively. In contrast, UrbanRefer targeted point-level generalized segmentation on terrestrial LiDAR road scenes, supporting both zero-target and multi-target mask prediction.

3D Referring Expression Segmentation. Compared with the well-developed 3D visual grounding task, 3D-RES has attracted increasing attention in recent years. Existing methods have mainly evolved from two-stage to single-stage, and more recently towards large model based frameworks. Early two-stage methods first generate candidate instances and then perform language-guided matching. For instance, TGNN [20] modeled instance-text relations with graph neural networks, while X-RefSeg3D [21] further enhanced cross-modal interaction by explicitly injecting language entities into scene graphs. To overcome the efficiency limitations of two-stage pipelines, single-stage methods directly fuse visual and linguistic features for end-to-end prediction. For example, 3D-STMN [2] introduced superpoint-text matching for efficient multimodal alignment, LESS [22] alleviated background interference through region regularization and contrastive learning, and RefMask3D [23] improved fine-grained alignment via a language primitive construction module. Other methods further enhanced spatial reasoning and feature completeness, such as RGSAN [24] using text-driven localization and weak supervision, and IPDN [25] using multi-view image features and task-oriented prompts. Beyond the traditional single-target setting, recent studies have extended 3D-RES to the more challenging 3D-GRES task [3], which requires handling single-target, multi-target, and zero-target scenarios. Meantime, inspired by the success of multimodal large models [26], recent works have begun to explore large model based 3D reasoning. Early methods such as SegPoint [27] and Reason3D [28] mainly focused on simple single-object scenarios, whereas MORE3D [29] and 3D-LLaVA [30] further extended this paradigm toward multi-object understanding and unified multi-task 3D vision-language learning.

Query Generation in 3D Scene Understanding. Query generation is a key step in language-guided 3D semantic segmentation. Traditional geometric FPS [31] works well in dense indoor scenes, but struggles in sparse and large-scale outdoor LiDAR environments, where it often oversamples background regions while missing semantically important yet geometrically sparse targets. Recent studies have therefore moved from geometric sampling toward adaptive query generation with multimodal priors. For example, OpenScene [32] introduced dense 2D vision-language features into 3D space through multi-view projection, OpenMask3D [33] improved query relevance via class-agnostic instance proposals and feature aggregation, and UniSeg3D [34] adopted learnable queries for dynamic generation across multiple segmentation tasks. Overall, current research trends emphasize guiding sampling toward information-rich regions by leveraging semantic saliency, feature density, or language priors, rather than relying solely on geometric uniformity. Motivated by this, we propose a text-weighted semantic FPS strategy with multi-granularity feature screening to generate more robust and informative queries.

3 Methodology

In this section, we first introduce the Cross-View Semantic Fusion (Sec. 3.1). Then, we describe the Semantic-Adaptive Query Generation (Sec. 3.2). Finally, we outline the Reliability-Aware Prompt Decoder (Sec. 3.3). An overview of our framework is shown in Fig. 1.

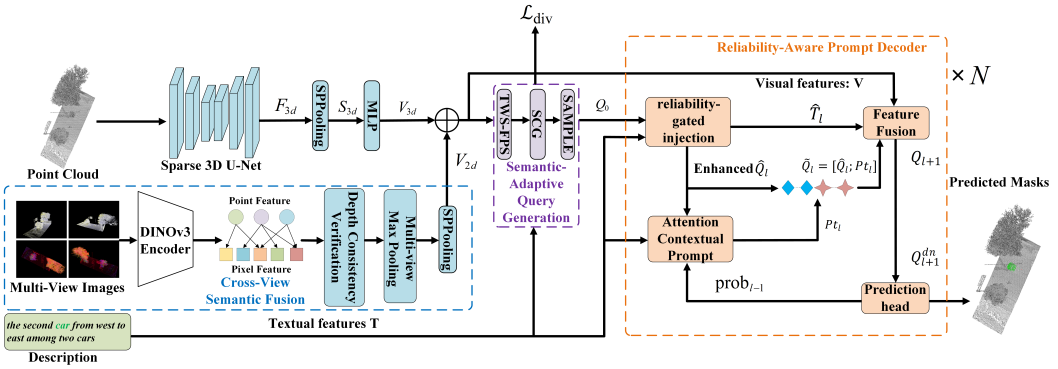

Figure 1: The overall framework of LaSA-Net. The input point clouds and multi-view images are first subjected to feature extraction and then aligned through cross-view semantic fusion. Subsequently, the resulting visual and textual features are fed into the Semantic-Adaptive Query Generation module to generate sparse text-relevant queries. Finally, the reliability-aware prompt decoder progressively performs prompt enhancement, multimodal fusion, and mask prediction, producing the final referring segmentation results.

3.1 Cross-View Semantic Fusion

After obtaining the textual features $T \in \mathbb{R}^{N \times d}$ and 3D visual features V_{3d} via feature extraction (see Appendix B for more details), given a 3D point cloud scene with N points and a set of RGB images captured from V different views, we first employ the pretrained DINOv3 (ViT-H/16+) [35] visual encoder to extract patch tokens from each image and reshape them into spatial feature maps. To achieve pixel-level alignment, we further upsample each spatial feature map to the original image resolution via bilinear interpolation, yielding dense pixel-wise 2D features $F_k^{\text{img}} \in \mathbb{R}^{C_I \times H \times W}$, where C_I denotes the feature dimension, and H and W denote the image height and width, respectively.

We then establish point-wise cross-modal correspondence by projecting the 3D point cloud onto the 2D image plane of each view. Notably, for indoor scenes, the camera intrinsic matrix K_k and extrinsic matrix T_k are directly obtained from sensor calibration. For outdoor scenes, they are derived from virtual camera poses and adaptive fields of view (FoV). Specifically, for a 3D point p_i under view k, we project it onto the image plane as follows:

$$u_i^k = f_x \frac{q_x}{q_z} + c_x, \qquad v_i^k = f_y \frac{q_y}{q_z} + c_y \tag{1}$$

where f_x and f_y denote the focal lengths along the horizontal and vertical directions, c_x and c_y denote the principal point coordinates, and q_x, q_y, and q_z denote the coordinates of the transformed 3D point in the camera coordinate system after applying the extrinsic transformation. To enforce geometric consistency under occlusion and projection noise, we introduce a depth-consistency check. For indoor scenes, the depth map is directly provided by the depth sensor, whereas for outdoor scenes, a z-buffer is constructed as its substitute. A 3D point is regarded as visible in view k only when its projection falls within the image boundary, its depth value is positive, and the absolute depth difference between its projected depth q_z and the corresponding depth value in D_k is smaller than a predefined threshold τ. The visibility mask is defined as follows:

$$\mathcal{V}_i^k = \mathbf{1}\left(0 \le u_i^k < W,\ 0 \le v_i^k < H,\ q_z > 0,\ \left|q_z - D_k(u_i^k, v_i^k)\right| < \tau\right) \tag{2}$$

where $\mathcal{V}_i^k \in \{0, 1\}$ denotes the visibility indicator of point p_i in the k-th view, D_k represents the corresponding depth map of this view, and τ is the depth-consistency threshold. Based on this criterion, we obtain the visibility mask $V_k \in \{0, 1\}^N$ and the normalized projection coordinates $(\hat{u}_i^k, \hat{v}_i^k) \in [-1, 1]^2$ for each view.

For all valid projections satisfying $\mathcal{V}_i^k = 1$, we use the normalized projection coordinates $(\hat{u}_i^k, \hat{v}_i^k)$ as sampling locations and perform bilinear interpolation on F_k^{img} to obtain the 2D semantic feature $f_i^k \in \mathbb{R}^{C_I}$ for point p_i in view k. We then aggregate the multi-view features by element-wise max pooling to form a robust point-wise global multi-view feature $f_i^{\text{mv}} \in \mathbb{R}^{C_I}$. Subsequently, these point-wise multi-view features are locally aggregated according to the superpoint assignment, producing

a superpoint-level multi-view semantic feature matrix $F_{2\mathrm{d}} \in \mathbb{R}^{N_s \times C_I}$, where N_s denotes the total number of superpoints in the scene. Finally, we map $F_{2\mathrm{d}}$ into the same embedding space as the 3D features through a MLP to obtain the 2D visual feature $V_{2\mathrm{d}} \in \mathbb{R}^{N_s \times d}$, which is added to the 3D visual feature $V_{3\mathrm{d}}$ to produce the final fused visual representation $V \in \mathbb{R}^{N_s \times d_v}$. Further details regarding the Cross-View Semantic Fusion can be found in Appendix C.

3.2 Semantic-Adaptive Query Generation

In the initial sampling stage, we design a Text-Weighted Semantic Farthest Point Sampling (TWS-FPS) strategy. Specifically, we introduce a lightweight semantic score predictor, where the superpoint-level visual features $V \in \mathbb{R}^{N_s \times d_v}$ and language features $T \in \mathbb{R}^{L \times d}$ are first projected into a shared d-dimensional space through linear layers, yielding $V' \in \mathbb{R}^{N_s \times d}$ and $T' \in \mathbb{R}^{L \times d}$. We then aggregate the variable-length text token sequence into a global text representation $T_{\mathrm{global}} \in \mathbb{R}^d$ by attention pooling using the following equation:

$$T_{\mathrm{global}} = \sum_{l=1}^{L} \alpha_l \cdot T'_l, \qquad \alpha_l = \frac{\exp\left(w^\top \tanh\left(W_a T'_l\right)\right)}{\sum_{l'} \exp\left(w^\top \tanh\left(W_a T'_{l'}\right)\right)} \tag{3}$$

where $W_a \in \mathbb{R}^{(d/2) \times d}$ and $w \in \mathbb{R}^{d/2}$ are learnable parameters. The resulting T_{global} is expanded along the spatial dimension and concatenated with each superpoint feature, and an MLP followed by a Sigmoid function is applied to predict a semantic relevance score $S_j \in [0, 1]$ for each superpoint.

The predicted score is then injected into the spatial distance metric in exponential form to construct a semantic-weighted distance as follows:

$$D_{\mathrm{sem}}(p_i, p_j) = \|p_i - p_j\|_2 \cdot \exp\left(\gamma \cdot S_j\right) \tag{4}$$

where p_i and p_j denote the superpoint coordinates, and γ is a hyperparameter that controls the strength of semantic guidance. In this way, candidates with higher semantic scores are assigned larger effective distances during FPS iterations and are therefore more likely to be selected as the farthest points. In addition, instead of random initialization, the sampling process starts from the superpoint with the highest semantic relevance score, i.e., $\arg\max_j S_j$, such that the most semantically relevant region is covered at the beginning of sampling. During training, TWS-FPS uses hard selection in the forward pass and a straight-through estimator with a soft-assignment temperature of 0.5 in the backward pass, allowing gradients to reach the semantic scores S_j.

After TWS-FPS, we sample seed points N_{seed} from the superpoints N_s, yielding the seed query feature $Q_{\mathrm{seed}} \in \mathbb{R}^{N_{\mathrm{seed}} \times d}$ and the corresponding 3D coordinates $p_{\mathrm{seed}} \in \mathbb{R}^{N_{\mathrm{seed}} \times 3}$. We then apply the SCG module [25] to the seed query feature Q_{seed} for local geometry-aware enhancement, producing the refined seed query $Q_{\mathrm{scg}} \in \mathbb{R}^{N_{\mathrm{seed}} \times d}$. This module exploits masked self-attention with a K-Nearest-Neighbor constraint in physical space to capture local spatial context around each seed point.

Subsequently, to enable more robust cross-modal sampling given complex referring expressions, we propose a Multi-Granularity Sampling (MGSampling) module. This module introduces parallel branches for word-level scoring and sentence-level global semantic matching, adaptively fusing these two granularity-specific scores via learnable weights. Specifically, given the enhanced seed queries $Q_{\mathrm{scg}} \in \mathbb{R}^{N_{\mathrm{seed}} \times d}$ and the language features $T \in \mathbb{R}^{L \times d}$, we compute fine-grained word-level matching scores through cross-attention to obtain R_i^{word}. Next, we aggregate the variable-length text sequence into a fixed-dimensional sentence embedding $T_{\mathrm{sent}} \in \mathbb{R}^d$ by masked mean pooling, and map each seed query into the same space using an unbiased linear projection $W_s \in \mathbb{R}^{d \times d}$. The sentence-level relevance score is then computed by cosine similarity followed by a ReLU nonlinearity. Finally, the two scores are fused into a unified correlation score:

$$R_i = R_i^{\mathrm{word}} + \sigma(\alpha_{\mathrm{raw}}) \cdot \mathrm{ReLU}\big(\cos(W_s Q_{\mathrm{scg},i},\, T_{\mathrm{sent}})\big) \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and α_{raw} is a learnable scalar initialized to -2.0, allowing the sentence-level semantic branch to gradually contribute during training. Since this design introduces only a single linear layer W_s, the additional parameter overhead is negligible. Based on the final correlation scores $R \in \mathbb{R}^{N_{\mathrm{seed}}}$, we perform hard Top-$K$ selection to retain the most relevant candidate points, which are then transformed by a three-layer MLP into the initial sparse query matrix $Q_0 \in \mathbb{R}^{m \times d}$ for the decoder. Further details regarding the Semantic-Adaptive Query Generation can be found in Appendix C.

3.3 Reliability-Aware Prompt Decoder

To mitigate the accumulated interference caused by unreliable language priors and noisy cross-modal cues during multi-layer query decoding, we propose a Reliability-Aware Prompt Decoder. Specifically, we first quantify the association between text tokens and queries through cross-attention. The text feature $T \in \mathbb{R}^{N_t \times d}$ is treated as the queries, while the sparse queries from the first layer, $Q_l \in \mathbb{R}^{m \times d}$, serve as the keys and values, yielding a cross-attention map $A_l \in \mathbb{R}^{N_t \times m}$. After transposing A_l, masking padding tokens, and renormalizing over the valid text tokens L_{valid}, we perform weighted aggregation over the text features to obtain the query-specific language prompt P_t. We then introduce a reliability gating mechanism based on the normalized Shannon entropy of the attention distribution. For the j-th query, the entropy-based reliability score is converted into a gating scalar r_j, which modulates residual prompt injection and produces the language-enhanced query $\hat{Q}_l$:

$$\hat{Q}_l = \text{LayerNorm}(Q_l + r \odot \text{MLP}_{\text{res}}(P_t)), \qquad r_j = \sigma\Big(\text{MLP}_r\Big(1 - H(A_l^j)\Big)\Big) \tag{6}$$

where $H(A_l^j) \in [0, 1]$ denotes the normalized entropy of the attention distribution for query j, $\sigma(\cdot)$ denotes the sigmoid function, $r \in \mathbb{R}^{m \times 1}$ is the query-wise gating vector formed by stacking all r_j, and $\odot$ denotes element-wise multiplication.

Finally, we use the foreground confidence predicted by the previous layer to perform threshold-based filtering over the query relevance score $\hat{S}_{c_l}$, Specifically, candidates with foreground probability below 0.75 are masked before the attention Softmax. We use the language-enhanced query $\hat{Q}_l$ to construct an attention-context prompt P_{t_l}. The prompt is concatenated with the query sequence along the query dimension to form the prompted query $\widetilde{Q}_l \in \mathbb{R}^{B \times 2m \times d}$:

$$\widetilde{Q}_l = \text{Concat}\Big(\hat{Q}_l, P_{t_l}\Big), \qquad P_{t_l} = \text{Softmax}\Big(\hat{S}_{c_l}\Big) \odot \hat{Q}_l \tag{7}$$

where the optimized $\hat{Q}_l \in \mathbb{R}^{B \times m \times d}$, refined by entropy-aware reliability gating, provides a semantically clean basis for constructing $P_{t_l} \in \mathbb{R}^{B \times m \times d}$.

Following the feature fusion strategy of MDIN [3], we use the prompted query $\widetilde{Q}_l$ to integrate the text feature $\hat{T}_l$ and the visual feature V, and update the queries under prompt guidance. The prompt tokens provide contextual information during subsequent self-attention, while only the first m core queries are fed into the prediction head. The prediction head produces the segmentation mask Mask_l and the target-instance probability Prob_l for layer l. The predicted Prob_l is further passed to the reliability-aware prompt module of layer $l + 1$ for threshold-based filtering.

4 Experiment

4.1 Benchmark Dataset

To support the proposed Outdoor 3D-GRES setting, we construct a large-scale outdoor benchmark, **UrbanRefer**, which extends our newly collected UrbanSeg dataset with referring expression annotations. The raw point clouds of UrbanSeg were collected in Simaqiao, Chengdu, China, and several European urban areas using *RIEGL VZ-600i*. More specifically, UrbanRefer contains 193 independent urban road scenes and $7{,}840$ language-scene pairs after filtering, with an average of 40.6 descriptions per scene. Using a fixed random seed, we split complete UrbanRefer scene tiles into training and validation sets at an 8:2 ratio, yielding 154 training scenes with 6,298 expressions and 39 validation scenes with 1,542 expressions. Experimental results obtained from UrbanRefer are evaluated on the validation set. In terms of sample composition, single-target (ST), multi-target (MT), and zero-target (ZT) account for 71.0%, 19.0%, and 10.0%, respectively. Notably, 61.5% of ST samples contain the same-category distractors, substantially increasing the difficulty of fine-grained spatial reasoning. Due to the complexity of urban road scenes, both target categories and referring expressions exhibit clear long-tail distributions, with trees, vehicles, and pedestrians dominating the dataset. Moreover, 85.0% of the descriptions involve topological, directional, or distance-based relations, highlighting the importance of relational cues in 3D-OGRES. Finally, to evaluate generalization across environments, we additionally conduct experiments on indoor 3D-GRES benchmarks. More details are provided in Appendix E.

Table 2: The outdoor 3D-GRES results on the UrbanRefer dataset.

Method	Acc@0.25 (%)				Acc@0.5 (%)				mIoU (%)
	ZT	ST	MT	All	ZT	ST	MT	All	
ReLA [36]	16.0/47.1	34.9/31.8	66.8	40.5	16.0/47.1	11.2/25.4	48.4	22.0	26.9
M3DRef-CLIP [37]	98/98	23.5/26.8	51.1	36.1	98/98	22.2/26.8	46.9	34.6	26.2
3D-STMN [2]	**100**/99	45.7/76.3	**92.4**	62.8	**100**/99	16.7/63.5	**67.5**	39.2	45.2
MDIN [3]	**100/100**	44.0/**87.9**	86.6	62.1	**100/100**	17.8/76.3	59.2	39.9	45.0
IPDN [25]	99/99	44.6/83.8	84.5	61.5	99/99	18.8/74.6	64.6	41.2	45.1
LaSA-Net (Ours)	98/99	**49.8**/81.5	85.2	**64.6**	98/99	**27.3/76.9**	61.7	**46.2**	**47.9**

Note: ZT, ST, and MT denote zero-target, single-target, and multi-target settings, respectively. For entries with "/", the left and right values indicate the cases with and without distractor objects, respectively.

Table 3: The 3D-GRES results on the Multi3DRefer dataset.

Method	Acc@0.25 (%)				Acc@0.5 (%)				mIoU (%)
	ZT	ST	MT	All	ZT	ST	MT	All	
ReLA [36]	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 [37]	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 [2]	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 [3]	**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 [25]	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
LaSA-Net (Ours)	46.6/**89.2**	**66.4/92.4**	78.6	**74.8**	46.6/**89.2**	**43.0/85.9**	**53.1**	**55.8**	**55.9**

Note: ZT, ST, and MT denote zero-target, single-target, and multi-target settings, respectively. For entries with "/", the left and right values indicate the cases with and without distractor objects, respectively.

4.2 Evaluation Criteria and Implementation Details

For evaluation, we adopt mean IoU (mIoU) and $\mathrm{acc@}k\mathrm{IoU}$, which measures the proportion of predictions whose IoU with the ground-truth mask exceeds a threshold $k \in \{0.25, 0.5\}$. For target-absent cases, a correct prediction is regarded as having an IoU of 1; otherwise, it is set to 0. In our experiments, the learning rate is scheduled using the PolyLR strategy, with an initial learning rate of 1×10^{-4} and a decay power of 4.0. The batch size is set to 4, and the number of queries m is set to 128. The decoder contains 6 layers. Our training objective includes a base loss $\mathcal{L}_b$, a probability loss $\mathcal{L}_p$, a contrastive learning loss $\mathcal{L}c$, and a diversity regularization loss $\mathcal{L}\mathrm{div}$, with corresponding weights of 1.0, 0.1, 0.1, and 0.1, respectively. Detailed loss definitions are provided in the Appendix D. All experiments are implemented in PyTorch 1.12.1 on an NVIDIA GeForce RTX 4090 GPU. Regarding computational efficiency, the online model contains 159.98M parameters, excluding the offline DINOv3 encoder. Training for 200 epochs takes approximately 14.5 hours with 20 GB peak GPU memory, while inference requires 11 ms per expression and 4.8 GB peak memory. Offline DINOv3 feature extraction for all 193 raw tiles takes about 20 minutes, including rendering and I/O.

4.3 Comparison with the State of the Art

To evaluate LaSA-Net, we compare it with 12 representative methods in three categories. The first category, including TGNN [20], InstanceRefer [10], 3D-STMN [2], X-RefSeg3D [21], and RefMask3D [23], focuses on conventional single-target 3D-RES with both two-stage and single-stage designs. The second category, including ReLA [36], M3DRef-CLIP [37], MDIN [3], and IPDN, addresses 3D-GRES and supports prediction from zero to multiple targets. The third category, consisting of SegPoint [27], Reason3D [28], and MCLN [38], is built on unified multimodal large language model (LLM) architectures and can handle complex user intentions while supporting multiple 3D perception tasks within a common framework. More details are provided in Appendix F.

As shown in Table 2, LaSA-Net achieves the best overall performance on the UrbanRefer dataset. Compared with earlier 3D-GRES methods, it improves $\mathrm{Acc@0.25}$ and $\mathrm{Acc@0.5}$ over ReLA by 24.1% and 24.2%, and exceeds M3DRef-CLIP by 21.7% in mIoU. Unlike prior methods that rely heavily on geometry-driven proposals or pre-computed instance priors, LaSA-Net uses semantic-adaptive query generation to guide sampling toward text-relevant regions, improving localization and

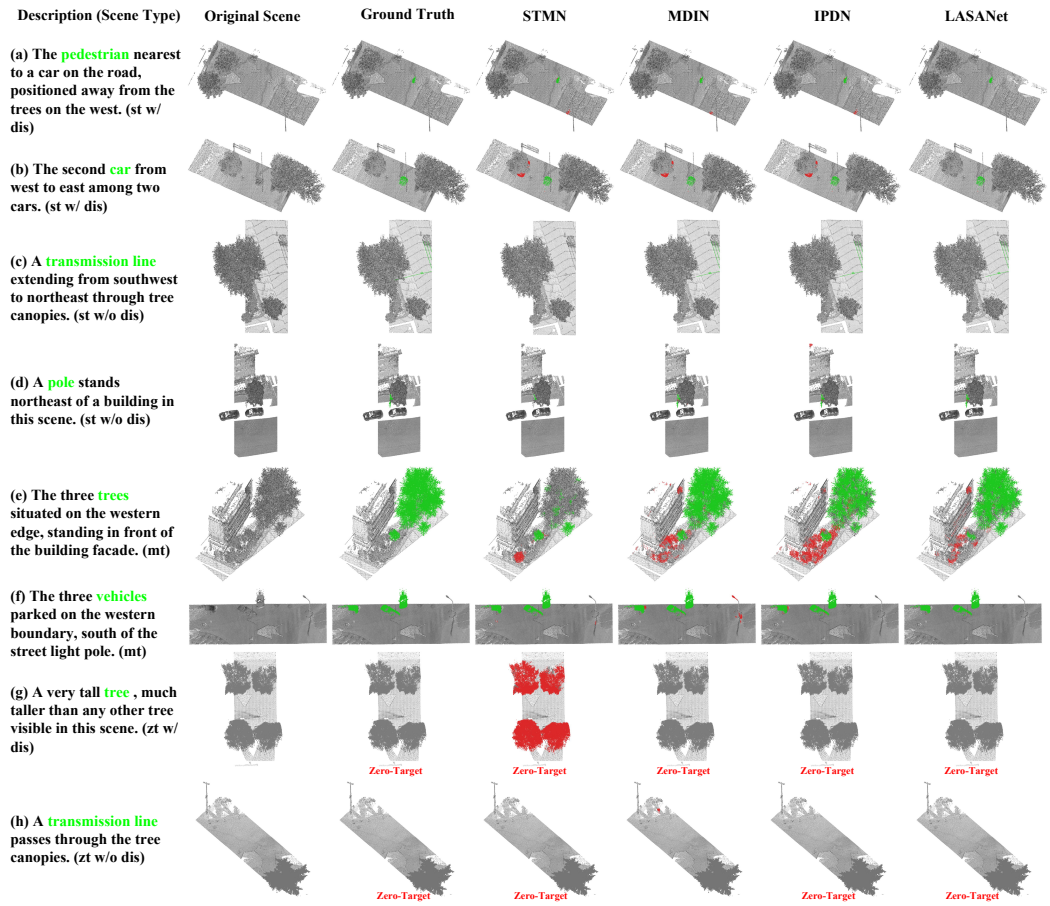

Figure 2: Qualitative comparison between SOTA methods and LaSA-Net on the UrbanRefer dataset. w/ dis and w/o dis indicate cases with and without distractors, respectively.

segmentation under complex backgrounds. Compared with recent methods using complex matching or image-enhanced decoding, LaSA-Net improves Acc@0.25 , Acc@0.5 , and mIoU by 1.8%, 5.0%, and 2.9%, respectively, over 3D-STMN, IPDN, and MDIN. This is because these methods remain sensitive to inaccurate hyperpoints, heavy background interference, or distracted global prompts, LaSA-Net introduces a reliability-aware decoder and cross-modal residual gating to suppress noisy language priors and maintain more stable query optimization.

Table 3 summarizes the quantitative comparison of representative 3D-GRES methods on the Multi3DRefer dataset. LaSA-Net achieves the best overall performance among all compared methods, reaching 74.8% in Acc@0.25 and 55.9% in mIoU. Compared with early methods, LaSA-Net shows clear advantages over ReLA and M3DRef-CLIP. ReLA relies on a two-stage proposal-and-matching pipeline and is limited by frozen instance proposals, making it vulnerable to incomplete candidates in cluttered scenes, especially for decentralized MT cases. M3DRef-CLIP directly transfers 2D vision–language priors in a zero-shot manner, but lacks adaptation to 3D spatial topology, leading to semantic ambiguity and instance confusion. Compared with recent methods such as 3D-STMN, MDIN, and IPDN, LaSA-Net still achieves the best overall results, surpassing IPDN by 3.3% in Acc@0.25, 5.8% in Acc@0.5, and 4.2% in mIoU. Although IPDN uses image-enhanced prompt decoding, its 2D compensation is limited by narrow indoor views and occlusions. In addition, 3D-STMN relies on rigid syntactic dependency priors, while MDIN is less sensitive to fine local boundaries. In contrast, LaSA-Net improves multi-target semantic consistency and fine-grained instance segmentation through semantic-adaptive query generation and multi-granularity sampling.

Fig. 2 shows that LaSA-Net consistently outperforms previous methods in multiple road scenarios. In the interference-heavy cases shown in Figs. 2(a) and 2(b), baseline methods such as 3D-STMN,

Table 4: The impact of core components.

Model	CSF	SAQG	RAPD	A@25	A@50	mIoU
I	×	×	×	62.8	39.9	45.4
II	✓	×	×	64.1	41.4	46.2
III	✓	✓	×	63.6	44.3	46.7
IV	✓	✓	✓	**64.6**	**46.2**	**47.9**

Note: A@25 and A@50 denote Acc@0.25 and Acc@0.5, respectively. All metrics are reported in %.

Table 5: The impact of different loss functions.

Model	$\mathcal{L}_{\mathbf{b}}$	$\mathcal{L}_{\mathbf{p}}$	$\mathcal{L}_{\mathbf{c}}$	$\mathcal{L}_{\mathbf{div}}$	A@25	A@50	mIoU
I	✓	×	×	×	55.2	31.5	37.8
II	✓	✓	×	×	63.3	40.5	45.9
III	✓	✓	✓	×	64.1	43.7	47.2
IV	✓	✓	✓	✓	**64.6**	**46.2**	**47.9**

Note: A@25 and A@50 denote Acc@0.25 and Acc@0.5, respectively. All metrics are reported in %.

Table 6: The impact of fine-grained components.

Model / variant	2D prior	Acc@0.25	Acc@0.50	mIoU
IPDN	CLIP	61.5	41.2	45.1
IPDN	DINOv3	60.1	42.5	45.6
LaSA-Net	DINOv3	**64.6**	46.2	**47.9**
w/o 2D branch	None	63.0	43.1	45.9
w/o TWS-FPS	DINOv3	63.6	45.1	47.0
w/o SCG	DINOv3	64.5	**46.4**	47.4
w/o Word-level Matching	DINOv3	63.9	42.6	46.3
w/o Sentence-level Matching	DINOv3	64.5	43.2	46.7

Note: All metrics are reported in %. CLIP denotes ViT-L/14, and DINOv3 denotes ViT-H/16+.

MDIN, and IPDN often produce localization errors and detect extra pedestrians or vehicles, reflecting their limited ability to reason over long-range spatial ambiguity and cluttered outdoor backgrounds. In the occluded cases shown in Figs. 2(c) and 2(d), 3D-STMN misses the transmission line, while MDIN and IPDN show limited accuracy on the pole. By contrast, LaSA-Net yields more complete structures and sharper boundaries, benefiting from DINOv3-based dense priors and precise depth-consistent 2D–3D fusion. In the multi-object scenarios shown in Fig. 2(e) and 2(f), existing methods still struggle with sparse multi-instance prediction under heavy background clutter. 3D-STMN misses one tree, while MDIN and IPDN produce only limited improvements and still suffer from multiple false predictions. This result reveals the weakness of conventional single-step prediction mechanisms, which treat all queries equally and thus cannot effectively focus on semantically relevant regions. In contrast, LaSA-Net successfully identifies all trees and cars. This advantage mainly comes from the semantic-adaptive query generation module, whose TWS-FPS and multi-granularity sampling improve both query initialization and global semantic consistency. In the zero-target scenarios shown in Fig. 2(g) and 2(h), methods such as 3D-STMN and MDIN are easily distracted by locally similar semantics, leading to noisy activations and false detections. This demonstrates the limited ability of conventional methods to reject invalid prompts in complex outdoor scenes. In contrast, LaSA-Net correctly predicts the zero-target condition and generates a clean empty mask. This robustness is mainly attributed to the reliability-aware decoder, which evaluates the reliability of language priors through attention entropy and suppresses noisy cross-modal cues with residual gating, thereby preventing error propagation when no valid target matches the description.

4.4 Ablation Study

Table 4 reports the ablation results of the three core components in LaSA-Net, namely Cross-View Semantic Fusion (CSF), Semantic-Adaptive Query Generation (SAQG), and the Reliability-Aware Prompt Decoder (RAPD). Model I, which removes all three modules, serves as the baseline and obtains 39.9% Acc@0.5 and 45.4% mIoU. Model II adds Cross-View Semantic Fusion on the basis of Model I, improving Acc@0.25, Acc@0.5, and mIoU from 62.8% to 64.1%, from 39.9% to 41.4%, and from 45.4% to 46.2%, respectively, showing the effectiveness of multi-view 2D–3D fusion in enhancing visual representation quality. Based on Model II, Model III further introduces Semantic-Adaptive Query Generation, which raises Acc@0.5 to 44.3% and mIoU to 46.7%, indicating its advantage in selecting text-relevant sparse queries for finer segmentation. Finally, Model IV incorporates the Reliability-Aware Prompt Decoder and achieves the best overall performance, with Acc@0.25, Acc@0.5, and mIoU reaching 64.6%, 46.2%, and 47.9%, respectively.

Table 5 presents the ablation results of different loss terms. Model I, which uses only the base loss $\mathcal{L}_b$, achieves relatively limited performance, with 31.5% Acc@0.5 and 37.8% mIoU. Model II further introduces the probability loss $\mathcal{L}_p$, which substantially improves Acc@0.5 to 40.5% and mIoU to 45.9%, respectively, and boosts zero-target performance from about 2% to 98%, highlighting the importance of explicit target-existence supervision. Model III adds the contrastive loss $\mathcal{L}_c$, further raising Acc@0.5 to 43.7% and mIoU to 47.2%, with especially clear gains in single-target scenarios, demonstrating the benefit of finer cross-modal semantic alignment. Finally, Model IV incorporates the diversity loss $\mathcal{L}_{\text{div}}$ and achieves the best overall performance, with Acc@0.5 and mIoU reaching 46.2% and 47.9%, respectively. The improvement is particularly notable in multi-target scenarios, where Acc@0.5 increases from 57.4% to 61.7%, showing that diversity regularization effectively alleviates query collapse and encourages different queries to focus on distinct targets.

Table 6 reports the controlled backbone comparison and fine-grained ablations of LaSA-Net on the UrbanRefer validation set. Replacing CLIP with DINOv3 in IPDN increases Acc@0.5 from 41.2% to 42.5% and mIoU from 45.1% to 45.6%, while Acc@0.25 slightly decreases from 61.5% to 60.1%, indicating that a stronger 2D encoder alone does not consistently improve referring segmentation. With the same DINOv3 prior, LaSA-Net achieves 64.6% Acc@0.25, 46.2% Acc@0.5, and 47.9% mIoU, substantially outperforming IPDN and demonstrating that the main improvement stems from the proposed architecture. Removing the 2D branch reduces the three metrics to 63.0%, 43.1%, and 45.9%, respectively, confirming the importance of dense 2D priors. Replacing TWS-FPS with standard FPS lowers the performance to 63.6%, 45.1%, and 47.0%, showing the effectiveness of text-guided seed selection. Removing word-level and sentence-level matching decreases Acc@0.5 to 42.6% and 43.2%, and mIoU to 46.3% and 46.7%, respectively, indicating that fine-grained lexical cues and global sentence semantics provide complementary guidance for target discrimination. Finally, removing SCG reduces mIoU to 47.4% while causing only minor fluctuations in localization accuracy, suggesting that spatial-context aggregation mainly improves mask quality.

5 Conclusion

In this work, we present LaSA-Net, a language-guided framework for generalized 3D referring expression segmentation in large-scale urban road scenes, together with UrbanRefer, the first benchmark for outdoor 3D-GRES. LaSA-Net addresses three central challenges in this setting, namely large-scale geospatial complexity in outdoor scenes, sparse-target perception under massive background interference, and context-induced query dilution caused by relation-dependent expressions. To this end, the framework integrates Cross-View Semantic Fusion for precise 2D–3D feature alignment, Semantic-Adaptive Query Generation for text-guided sparse query selection, and a Reliability-Aware Prompt Decoder for adaptive language prompt injection and robust iterative refinement. Extensive experiments on ScanRefer, Multi3DRefer, and UrbanRefer show that LaSA-Net consistently outperforms previous SOTA methods, achieving mIoU gains of 2.3%, 4.2%, and 2.8%, respectively. These results demonstrated strong robustness under complex outdoor spatial relations and highly sparse target distributions, and established an effective solution for spatial semantic understanding in applications such as urban digital twins and autonomous driving.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 42371447, the Research on Key Technologies and Industrial Application of Satellite-Ground Collaborative Intelligent Remote Sensing Services under Grant 26GF2700100, and the ECNU Academic Innovation Promotion Program for Excellent Doctoral Students under Grant YBNLTS2026013.

References

[1] Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, and Liujuan Cao. 3d-dres: Detailed 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2966–2974, 2026.

[2] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5940–5948, March 2024.

[3] 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, Melbourne VIC Australia, October 2024. ACM.

[4] Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *arXiv preprint arXiv:2508.00265*, 2025.

[5] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023.

[6] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form Description Guided 3D Visual Graph Network for Object Grounding in Point Cloud. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3702–3711, Montreal, QC, Canada, October 2021. IEEE.

[7] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. SAT: 2D Semantics Assisted Training for 3D Visual Grounding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1836–1846, Montreal, QC, Canada, October 2021. IEEE.

[8] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation Modeling for Visual Grounding on Point Clouds. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2908–2917, Montreal, QC, Canada, October 2021. IEEE.

[9] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. TransRefer3D: Entity-and-Relation Aware Transformer for Fine-Grained 3D Visual Grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, Virtual Event China, October 2021. ACM.

[10] 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1771–1780, Montreal, QC, Canada, October 2021. IEEE.

[11] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. Distilling Coarse-to-Fine Semantic Matching Knowledge for Weakly Supervised 3D Visual Grounding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2662–2671, Paris, France, October 2023. IEEE.

[12] Yang Liu, Daizong Liu, and Wei Hu. Joint top-down and bottom-up frameworks for 3d visual grounding. In *International Conference on Pattern Recognition*, pages 249–264. Springer, 2024.

[13] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16433–16442, New Orleans, LA, USA, June 2022. IEEE.

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

[15] Atharv Mahesh Mane, Dulanga Weerakoon, Vigneshwaran Subbaraju, Sougata Sen, Sanjay E Sarma, and Archan Misra. Ges3vig: Incorporating pointing gestures into language-based 3d visual grounding for embodied reference understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9017–9026, 2025.

[16] Fuhao Li, Huan Jin, Bin Gao, Liaoyuan Fan, Lihui Jiang, and Long Zeng. Nugrounding: A multi-view 3d visual grounding framework in autonomous driving. *arXiv preprint arXiv:2503.22436*, 2025.

[17] Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, Youquan Liu, Dongyue Lu, Liang Pan, Lingdong Kong, Junwei Liang, and Ziwei Liu. 3eed: Ground everything everywhere in 3d. *arXiv preprint arXiv:2511.01755*, 2025.

[18] Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. Cityrefer: Geography-aware 3d visual grounding dataset on city-scale point cloud data. *arXiv preprint arXiv:2310.18773*, 2023.

[19] Jinpeng Li, Haiping Wang, Yuan Liu, Zhiyang Dou, Yuexin Ma, Sibei Yang, Yuan Li, Wenping Wang, Zhen Dong, Bisheng Yang, et al. Cityanchor: City-scale 3d visual grounding with multi-modality llms. In *International Conference on Learning Representations*, volume 2025, pages 39415–39438, 2025.

[20] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-Guided Graph Neural Networks for Referring 3D Instance Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1610–1618, May 2021.

[21] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured Cross-Modal Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4551–4559, March 2024.

[22] Xuexun Liu, Xiaoxu Xu, Jinlong Li, Qiudan Zhang, Xu Wang, Nicu Sebe, and Lin Ma. Less: Label-efficient and single-stage referring 3d segmentation. *Advances in Neural Information Processing Systems*, 37:11164–11185, 2024.

[23] 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.

[24] Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. 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.

[25] 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*, volume 39, pages 2132–2140, 2025.

[26] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[27] 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.

[28] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In *2025 International Conference on 3D Vision (3DV)*, pages 1177–1186. IEEE, 2025.

[29] Xueying Jiang, Lewei Lu, Ling Shao, and Shijian Lu. Multimodal 3d reasoning segmentation with complex scenes. *arXiv preprint arXiv:2411.13927*, 2024.

[30] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3772–3782, 2025.

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[32] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023.

[33] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

[34] Wei Xu, Chunsheng Shi, Sifan Tu, Xin Zhou, Dingkang Liang, and Xiang Bai. A unified framework for 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:59468–59490, 2024.

[35] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[36] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized Referring Expression Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601, Vancouver, BC, Canada, June 2023. IEEE.

[37] Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3DRefer: Grounding Text Description to Multiple 3D Objects. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15179–15179, Paris, France, October 2023. IEEE.

[38] 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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15104, pages 381–398. Springer Nature Switzerland, Cham, 2025. Series Title: Lecture Notes in Computer Science.

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

[40] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks, November 2017. arXiv:1711.10275.

[41] 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.

[42] 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.

Appendix

Contents

A Problem Statement **15**

B Feature Extraction **15**

C Module Illustration **15**

D Loss Function **15**

E Descriptions of Datasets **17**
E.1 Outdoor Datasets 17
E.2 Indoor Datasets 20

F Comparison on Indoor Datasets **21**
F.1 ScanRefer Dataset 21
F.2 Multi3DRefer Dataset 23

G Limitations **24**

H Broader Impact **24**

A Problem Statement

We consider generalized 3D referring expression segmentation in large-scale outdoor environments, which fundamentally differ from indoor settings in spatial scale, relational structure, and target distribution. In particular, outdoor point clouds exhibit extended spatial extent, require long-range geospatial reasoning, and contain sparse targets embedded within dense same-category distractors. Formally, a point cloud scene is represented as $P = \{\mathbf{p}_i\}_{i=1}^{N}$, where each point $\mathbf{p}_i \in \mathbb{R}^{3+d}$ encodes its 3D coordinates and optional appearance attributes. Let E denote a natural-language referring expression. Following the 3D-GRES formulation, we allow the referred set to contain zero, one, or multiple objects. The goal is to predict a binary point-wise mask $M \in \{0,1\}^N$, where each element indicates whether the corresponding point belongs to the object set referred to by E. Extending this formulation to the proposed outdoor setting, the problem can be defined as learning a mapping as:

$$M^* = \arg \max_{M \in \{0,1\}^N} \; p_\theta(M \mid P, E) \tag{8}$$

where $p_\theta(\cdot)$ denotes a parameterized model that measures the likelihood of the mask M given the input point cloud P and referring expression E, and M^* is the optimal prediction.

B Feature Extraction

Textual Feature Extraction. Given a textual description for the target object, we utilize a pre-trained RoBERTa [39] to extract word-level embeddings $E \in \mathbb{R}^{N_t \times C_t}$, where N_t denotes the number of tokens and C_t indicates the dimensionality of each embedding. To obtain a unified feature dimension d in the decoder, we transform E into textual features $T \in \mathbb{R}^{N_t \times d}$ via a linear projection:

$$T = EW_t \tag{9}$$

where $W_t \in \mathbb{R}^{C_t \times d}$ denotes the learnable projection matrix.

Visual Feature Extraction. Consider a point cloud scene $P \in \mathbb{R}^{N_p \times (3+f)}$ with N_p points, each point consists of a 3D coordinate and an auxiliary feature vector of dimension f, such as RGB values or normal vectors. We first apply a sparse 3D U-Net [40] to extract point-wise features $F_{3d} \in \mathbb{R}^{N_p \times C_p}$, where C_p denotes the feature dimensionality. Next, we generate N_s superpoints $\{SP_i\}_{i=1}^{N_s}$ from the original point cloud and perform superpoint pooling on F_{3d} to obtain 3D superpoint features $S_{3d} \in \mathbb{R}^{N_s \times C_p}$. Finally, a MLP projects these features into dimension d, producing the 3D visual features $V_{3d} \in \mathbb{R}^{N_s \times d}$ using the following equation:

$$V_{3d} = \mathrm{MLP}\big(\mathrm{SPPool}(F_{3d})\big) \tag{10}$$

where $\mathrm{MLP}(\cdot)$ denotes a learnable multi-layer perceptron, and $\mathrm{SPPool}(\cdot)$ denotes the superpoint pooling operation.

C Module Illustration

The detailed module illustrations of Cross-View Semantic Fusion and Semantic-Adaptive Query Generation are shown in Fig. 3 and Fig. 4, respectively.

D Loss Function

The overall loss function of the proposed method consists of four parts. The first part is the base loss $\mathcal{L}_b$, which is applied to the queries matched with target instances and includes the binary cross-entropy loss and Dice loss for mask prediction, as well as the mean squared error loss for IoU score regression. The second part is the probability loss $\mathcal{L}_p$, which supervises the foreground probability Prob of each query. The third part is the contrastive learning loss $\mathcal{L}_c$, which is used to align the text features with their corresponding queries. The fourth part is the Margin-Focal diversity regularization loss $\mathcal{L}_{\text{div}}$ for feature collapse prevention during sampling. For a feature pair (q_i, q_j) in the candidate set $\mathcal{P}$, with corresponding coordinates (p_i, p_j), the diversity term is defined as follows:

$$s_{ij} = \max(\cos(q_i, q_j) - \mu, 0)\,, \qquad \mathcal{L}_{\text{div}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)} \left(\frac{s_{ij}}{1-\mu}\right)^{\alpha} \cdot s_{ij} \cdot \exp\left(-\frac{\|p_i - p_j\|_2^2}{\sigma}\right) \tag{11}$$

where μ denotes the margin threshold, α denotes the focal weighting factor, and σ controls the spatial distance decay. All loss terms are combined through weighted summation as follows:

$$\mathcal{L} = \lambda_b \mathcal{L}_b + \lambda_p \mathcal{L}_p + \lambda_c \mathcal{L}_c + \lambda_{\text{div}} \mathcal{L}_{\text{div}} \tag{12}$$

where λ_b, λ_p, λ_c, and λ_{div} are balancing coefficients for the corresponding loss terms.

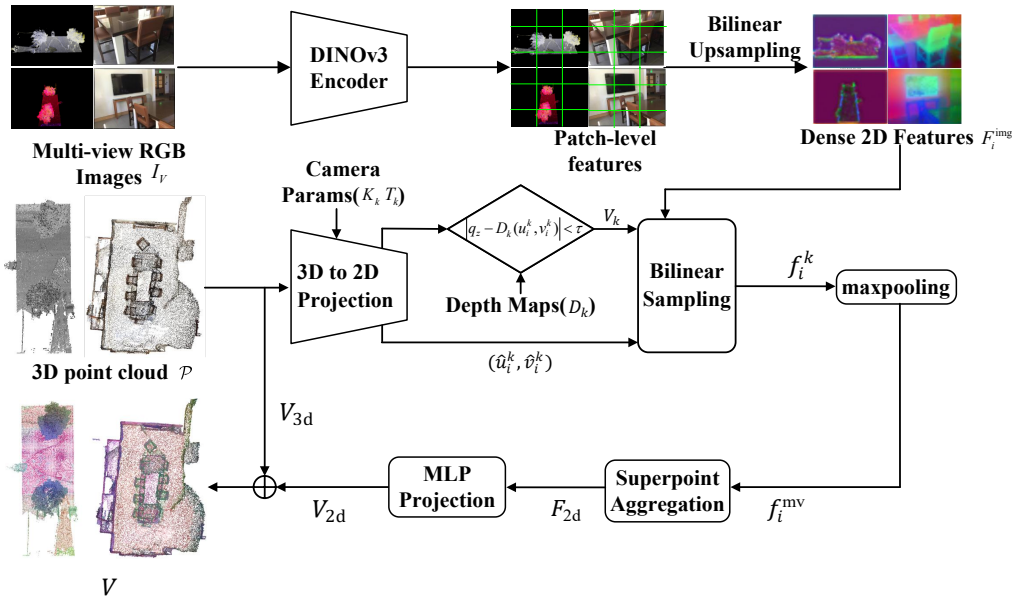

Figure 3: Illustration of the Cross-View Semantic Fusion.

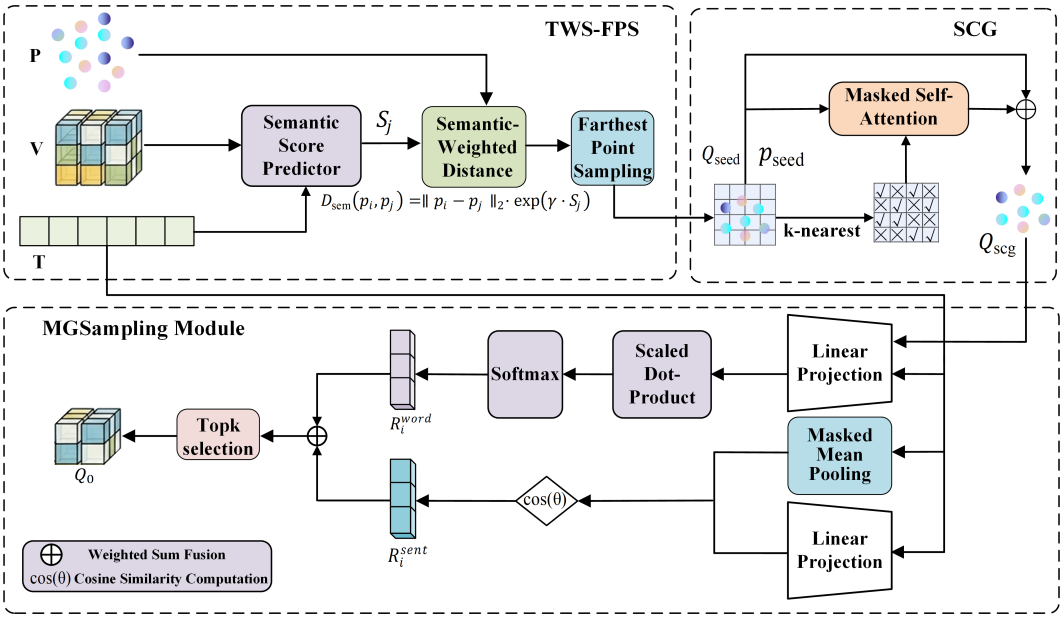

Figure 4: Illustration of the Semantic-Adaptive Query Generation.

E Descriptions of Datasets

E.1 Outdoor Datasets

E.1.1 UrbanSeg Construction

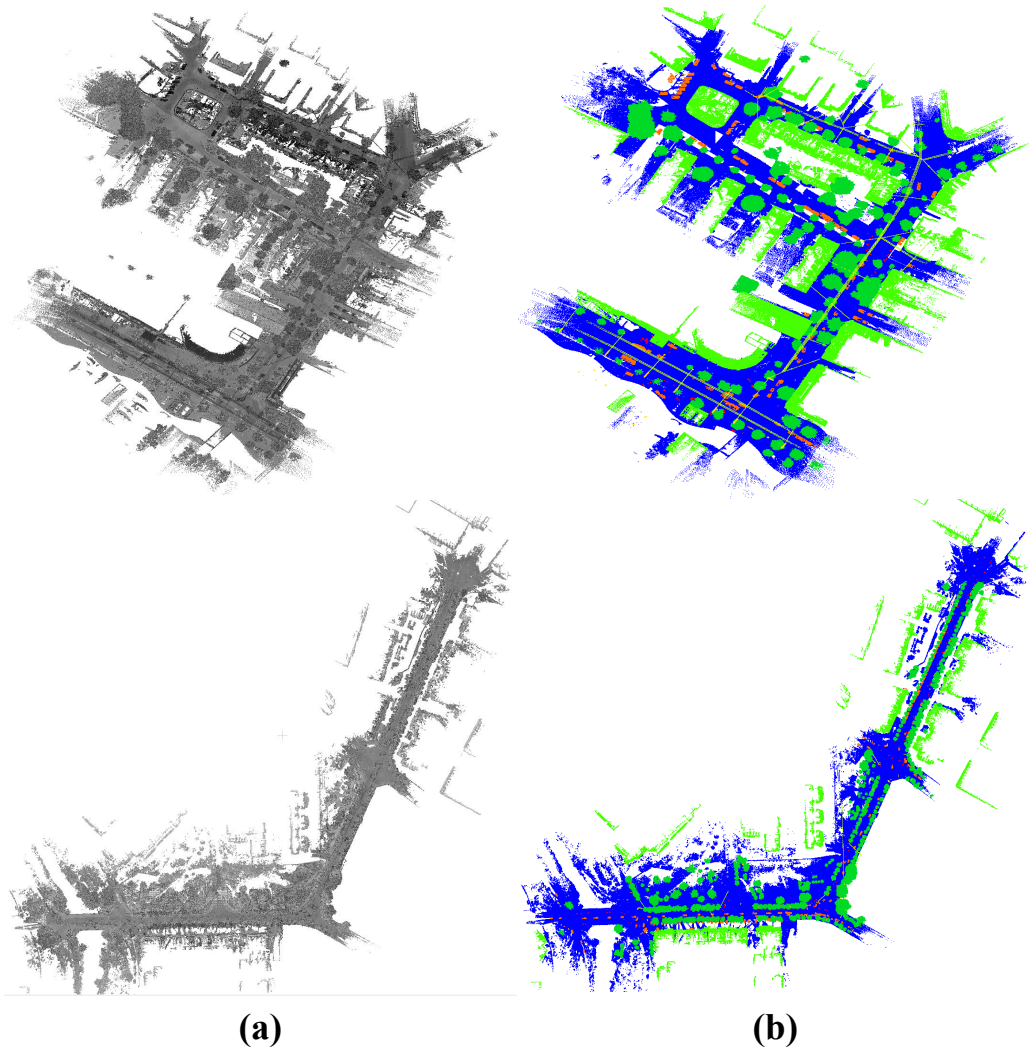

Figure 5: Overview of the UrbanSeg dataset, which is shown with (a) natural color (RGB) and (b) instance labels.

As shown in Fig. 5, following the annotation protocol of the Toronto-3D benchmark, we first annotated the raw point clouds with point-level semantic labels. Each point was assigned to one of nine categories: *others* (e.g., trash cans and fences), *road*, *trees*, *buildings*, *transmission lines*, *poles* (including traffic-light poles and street-light poles), *cars*, *pedestrians* (including cyclists), and *low vegetation* (e.g., lawns and shrubs). All annotations were manually completed in CloudCompare software and verified through cross-validation.

Based on the semantic labels, we further performed instance-level annotation. To avoid incorrect merging across categories, the point cloud was first separated by semantic class. Connected-component labeling was then applied to each class-specific subset with tuned parameters, followed by manual refinement to ensure spatial separation and geometric completeness of each instance. Afterward, all annotated instances were merged back into a unified point cloud, and the source fragment of each object was recorded as a local instance identifier. To ensure global uniqueness across the dataset, we

defined the global instance identifier as follows:

$$\mathrm{InstanceID}_{\mathrm{Global}} = 10000 \times \mathrm{SemanticID} + \mathrm{InstanceID}_{\mathrm{Local}}. \tag{13}$$

This formulation uniquely indexes each instance while preserving its semantic category information.

In addition, the originally acquired data consist of several large-scale panoramic point cloud datasets with substantial data volume and long spatial extent, which are difficult to process directly in deep learning frameworks due to GPU memory limitations and computational bottlenecks. Therefore, the raw point clouds are further partitioned into small scenes for subsequent data loading, network training, and inference. Specifically, we first extract the road-surface points and apply Principal Component Analysis (PCA) to estimate the principal direction of the road. Let the road-surface point set be denoted as $P_{\mathrm{road}} = \{p_i \mid p_i \in \mathbb{R}^3\}$. PCA is then applied to compute the dominant direction vector $\mathbf{d} \in \mathbb{R}^3$, which represents the global longitudinal direction of the road. Based on this direction, the scene is segmented every $10\,\mathrm{m}$ along $\mathbf{d}$, while an overlap of $3\,\mathrm{m}$ is preserved between adjacent sub-scenes to reduce boundary discontinuity.

To suppress interference from distant background, reduce redundant computation, and make the model focus on the core traffic area of each scene, we further constrain the retained spatial width. Specifically, only the region satisfying $Y \in [y_{\mathrm{center}} - 15\,\mathrm{m},\, y_{\mathrm{center}} + 15\,\mathrm{m}]$ is preserved, where y_{center} denotes the road centerline in the lateral direction. In practice, this setting corresponds to a road body of approximately $10\,\mathrm{m}$ in width together with outward extension on both sides to retain sidewalks and building facades, yielding a total effective cropping width of $30\,\mathrm{m}$. Background points outside this range, such as distant vegetation and buildings, are directly removed.

To avoid truncating objects at scene boundaries, we further adopt an integrity-preserving boundary assignment strategy based on instance centroids and bounding boxes. For an instance I_k, let its centroid be denoted as $\mathbf{c}_k \in \mathbb{R}^3$ and its bounding box as B_k. According to the spatial location of $\mathbf{c}_k$ relative to the cropping boundary, an instance intersecting the segmentation line is either fully retained or fully discarded, rather than being partially split across adjacent sub-scenes. This strategy ensures the geometric completeness of cross-boundary objects and avoids abrupt instance truncation caused by naive spatial clipping. Finally, after cropping, filtering, and quality screening, we obtain 193 high-quality outdoor scenes, which are used for the dataset construction of UrbanRefer.

E.1.2 Referring Expression Generation for UrbanRefer

As shown in Fig. 6, inspired by Multi3DRefer [37] and Sr3D [41], we adopt a pipeline that first generates spatial-relation descriptions with rule-based templates and then refines them using a LLM, namely GPT-5.4. During data preprocessing, we extract geometric priors of all point cloud instances, including centroids, bounding boxes, and absolute heights, and removes noisy instances containing fewer than 10 points. Based on semantic mapping rules, all scene elements are further organized into a target–anchor topological network. Specifically, the valid target set, denoted as $\mathcal{T}$, includes *trees*, *power lines*, *utility poles*, *vehicles*, and *pedestrians*, while the spatial reference anchor set, denoted as $\mathcal{A}$, includes *roads*, *buildings*, and *low vegetation*. To accommodate the task characteristics of 3D-GRES, we define three description-generation settings, namely zero-target (ZT), single-target (ST), and multi-target (MT).

Zero-Target descriptions. To improve the model's capability for answer rejection and enhance robustness under ambiguous or misleading language input, we further construct two categories of ZT descriptions. The total number of such descriptions is strictly controlled to remain within 10% of the number of valid instances. The first type is *category absence*, where category-level descriptions that are logically plausible but absent from the current scene are generated. The second type is *attribute or spatial mismatch*, where the spatial orientation or physical attributes of existing instances are deliberately perturbed to create misleading references. These ZT descriptions provide hard negative samples for training and evaluation, thereby improving the model's ability to distinguish between valid targets and non-existent referents.

Single-Target descriptions. For the ST setting, we design category-specific templates for different target types. These templates cover several complementary spatial and geometric reasoning strategies. First, for *azimuth topology*, the azimuth of a target with respect to its anchor is computed under an eight-direction compass system, so that expressions such as northeast, southwest, or east of a building can be generated in a structured way. Second, for *axis ordinal reasoning*, instances of the same

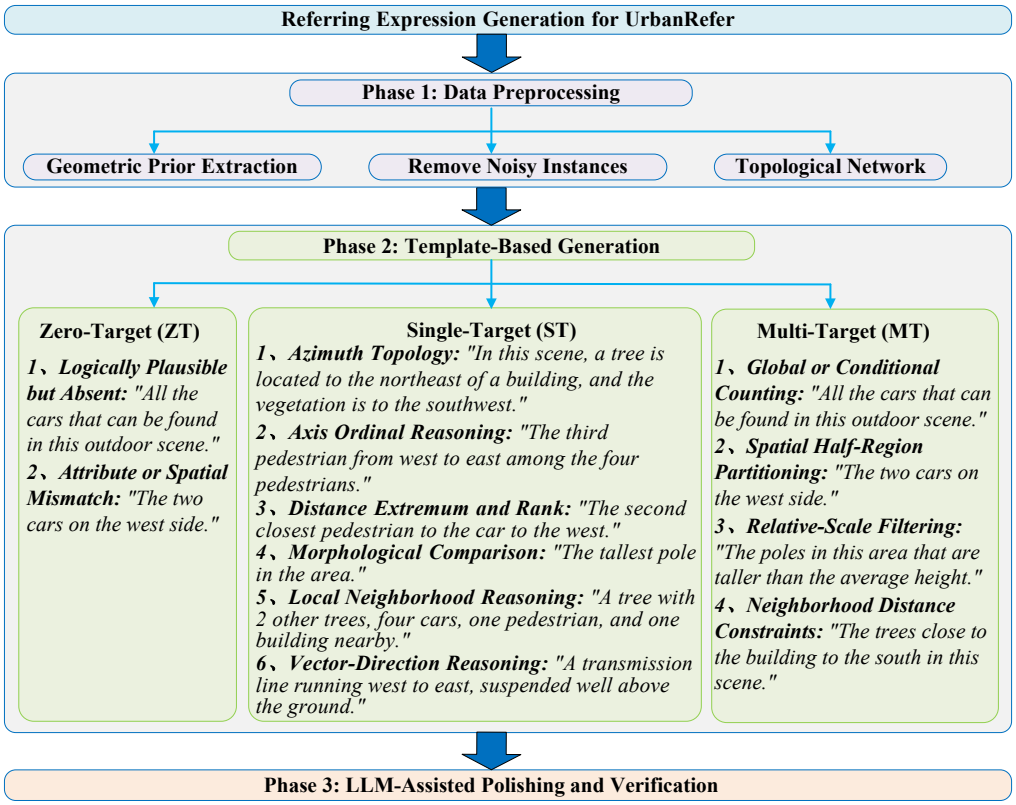

Figure 6: Pipeline of the referring expression generation for UrbanRefer.

semantic type are sorted along the world-coordinate axes, enabling ordinal descriptions such as the leftmost, rightmost, frontmost, or rearmost target. Third, for *distance extremum and rank*, the absolute Euclidean distance between the target and a designated anchor is verified, and then either an extremum description, such as the closest one, or an ordered description, such as the second closest one, is generated accordingly. Fourth, for *morphological comparison*, significant geometric differences are converted into descriptions based on height deviation, such as the tallest or the shortest instance in a local group. Fifth, for *local neighborhood reasoning*, the contextual entity distribution within a radius of $15\,\mathrm{m}$ centered at the target is counted and summarized to form descriptions involving nearby roads, buildings, poles, or vegetation. Sixth, for *vector-direction reasoning*, the spatial extension direction of line-structured objects, such as power lines, is extracted and transformed into descriptive texts. In addition to the core template content, we further improve linguistic diversity and naturalness by adding prefixes and suffixes, and then perform description deduplication and truncation. After this process, at most six high-quality descriptions are retained for each valid instance.

Multi-Target descriptions. For the MT setting, we generate expressions that jointly refer to a subset of instances of the same semantic type. To achieve this, a feature-filtering strategy is introduced to identify target groups with shared spatial or geometric characteristics. The MT selection process includes several operations. First, *global or conditional counting* is used to describe subsets under scene-level or condition-specific constraints. Second, *spatial half-region partitioning* is performed according to the coordinate median of candidate instances, so that groups located in one half of the scene can be described. Third, *relative-scale filtering* is used to select subsets with distinctive geometric properties, such as relatively taller or shorter instances. Fourth, *neighborhood-distance constraints* are applied to identify groups of targets that satisfy specific proximity relationships. Through these strategies, the generated descriptions can accurately characterize multiple targets that share common features and thus conform to the requirements of generalized referring segmentation.

Overall, the above pipeline allows UrbanRefer to support diverse referring settings ranging from single-instance localization to multi-instance grouping and zero-target rejection, while maintaining high consistency between structured geometric priors and natural-language expressions.

LLM-assisted polishing and verification. To improve the naturalness, fluency, and correctness of the generated expressions, we adopt a unified quality-control pipeline consisting of LLM-assisted polishing and manual verification. In the polishing stage, GPT-5.4 is used to refine template-generated descriptions by optimizing syntactic structure, removing redundant modifiers, and improving fluency, while strictly preserving the original object category, spatial relationship, and physical attribute information. To enhance linguistic diversity and avoid overfitting to a single rewriting pattern, the polishing process is implemented in batches of 20 samples, with five groups of differentiated prompts used alternately, and an exponential-backoff retry mechanism is further introduced to ensure stable processing. Quantitatively, this stage reduces the average description length from 14.5 words to 11.4 words, while lowering the occurrence frequency of high-frequency templated phrases such as "there is a" from 16.4% to 1.5%, indicating that the polished descriptions become more concise and natural. In the subsequent verification stage, well-trained annotators are assigned to identify the referred objects or object set in the corresponding 3D scene, remove descriptions that are invalid or ambiguous, and manually correct minor spelling, wording, and grammatical issues when necessary. Consequently, $1{,}833$ invalid descriptions are filtered out from 9,673 candidates, leaving 7,840 final descriptions, and 2,239 descriptions are manually corrected, further improving the reliability and usability of UrbanRefer for subsequent training and evaluation.

E.1.3 Multi-view Pseudo-color Image Generation

For indoor scenes, given a 3D point cloud $P \in \mathbb{R}^{N\times 3}$ with N points, together with RGB images $\{I_k\}_{k=1}^{V}$ and corresponding depth maps $\{D_k\}_{k=1}^{V}$ captured from V real sensor viewpoints, such multimodal observations can be directly fed into the feature extraction module. In contrast, large-scale outdoor urban scenes usually provide only LiDAR point clouds, without aligned multi-view RGB images or depth observations. To bridge this visual modality gap, we design a reconstruction pipeline based on LiDAR physical-property pseudo-color encoding and virtual-camera rendering.

Specifically, we first extract five key physical attributes from the raw LiDAR observations, including reflection intensity, calibrated reflectance, signal amplitude, return count, and waveform deviation, and then transfer them to the downsampled target point clouds through KD tree based coordinate matching. To enrich semantic cues, these attributes are further encoded into two complementary pseudo-color RGB representations. The first is a *material view*, in which intensity, reflectance, and amplitude are mapped into three RGB channels after quantile normalization for contrast enhancement. The second is a *structural view*, in which elevation, return count, and waveform deviation are encoded into another RGB image to emphasize geometric and return-pattern variations.

Based on the 3D scene bounding box, we further place V' virtual cameras automatically to cover top-down views, 45° oblique views, and 15° low-angle side views. An adaptive FoV strategy is adopted to maintain an image occupancy ratio of approximately 80%, so that the projected scene content can make effective use of the rendered image area. We then employ Open3D offscreen rendering to synthesize the pseudo-color images. Examples of the synthesized multi-view pseudo-color images are shown in Fig. 7. As a result, this process produces $2V'$ synthesized multi-view images $\{I_{s,k}^{\mathrm{pseudo}}\}$, while simultaneously providing the intrinsic and extrinsic parameters of all virtual cameras together with depth information derived from the Z-buffer. Accordingly, LiDAR-only outdoor scenes are transformed into multimodal multi-view observations that are structurally equivalent to those commonly available in indoor settings, thereby enabling subsequent cross-modal feature extraction and geometric projection under a unified formulation.

E.2 Indoor Datasets

ScanRefer. We use the ScanRefer dataset [42] to evaluate our method on the 3D-RES task. ScanRefer consists of 51,583 natural language expressions, covering 11,046 referred objects across 800 ScanNet scenes. In addition, we adopt mIoU, Acc@0.25, and Acc@0.5 as the evaluation metrics.

Multi3DRefer. We further use the Multi3DRefer dataset [37] to evaluate the performance of our model on the 3D-GRES task. Different from conventional 3D-RES, 3D-GRES allows the number of

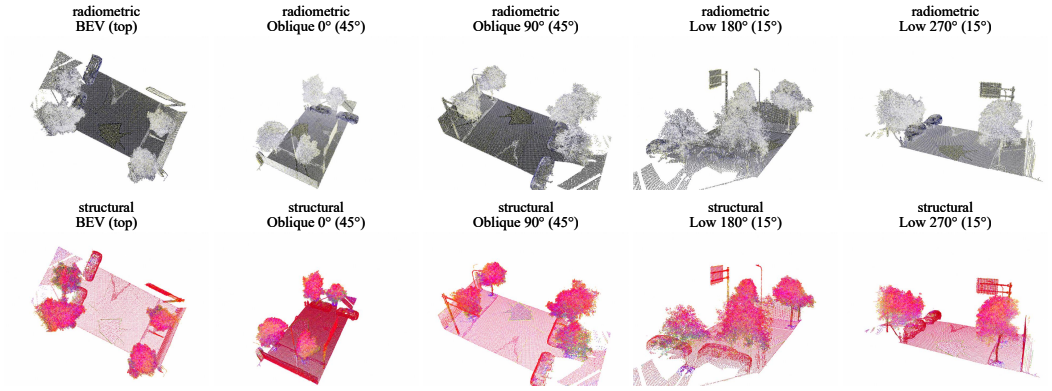

Figure 7: Multi-view pseudo-color renderings of representative UrbanRefer scenes.

Table 7: The 3D-RES results on ScanRefer.

Method	Reference	Unique (∼19%)		Multiple (∼81%)		Overall @0.25(%)	Overall @0.5(%)	mIoU(%)
		@0.25	@0.5	@0.25	@0.5			
TGNN [20]	AAAI'21	-	-	-	-	37.5	31.4	27.8
TGNN† [20]	AAAI'21	69.3	57.8	31.2	26.6	38.6	32.7	28.8
InstanceRefer† [10]	ICCV'21	81.6	72.2	29.4	23.5	40.2	33.5	30.6
X-RefSeg3D [21]	AAAI'24	-	-	-	-	40.3	33.8	29.9
3D-STMN [2]	AAAI'24	89.3	84.0	46.2	29.2	54.6	39.8	39.5
SegPoint [27]	ECCV'24	-	-	-	-	-	-	41.7
Reanson3D [28]	3DV'25	88.4	84.2	50.5	31.7	57.9	41.9	42.0
RefMask3D [23]	ACMMM'24	89.6	84.7	48.1	40.8	55.9	49.2	44.9
MCLN [38]	ECCV2024	89.6	78.2	53.3	45.9	58.7	50.7	44.7
MDIN [3]	ACMMM'24	91.0	87.2	50.1	44.9	58.0	53.1	48.3
IPDN [25]	AAAI'25	91.5	88.0	53.1	47.0	60.6	54.9	50.2
LaSA-Net (Ours)		**94.4**	**90.1**	**54.6**	**49.6**	**62.3**	**57.5**	**52.5**

Note: † The mIoU and accuracy are reevaluated by ourself. "Unique" denotes samples with a single instance of the referred category in the scene, while "Multiple" denotes samples with multiple such instances.

targets referred to by a language expression to be arbitrary, including zero, one, or multiple objects. Multi3DRefer contains 61,926 language descriptions in total, among which 51,583 descriptions are directly inherited from ScanRefer. Specifically, 6,688 descriptions correspond to zero-target cases, 13,178 descriptions correspond to multi-target cases, and the remaining descriptions correspond to single-target cases. We adopt the same evaluation metrics as those used for ScanRefer. For zero-target samples, the mIoU is set to 1 if the model correctly predicts that no target object is present, and 0 otherwise.

F Comparison on Indoor Datasets

F.1 ScanRefer Dataset

Table 7 reports the quantitative comparison on the ScanRefer dataset. LaSA-Net achieves the best overall performance, reaching 62.3% in $\mathrm{Acc@0.25}$, 57.5% in $\mathrm{Acc@0.5}$, and 52.5% in mIoU. Compared with early two-stage or candidate-matching methods, such as TGNN, InstanceRefer, and X-RefSeg3D, LaSA-Net shows clear advantages because these methods remain fundamentally constrained by the quality of pre-generated candidates or intermediate clustering, making them prone to cumulative errors and limited recovery of missed geometries. By contrast, LaSA-Net removes the inefficient generate-then-match pipeline and directly performs text-guided end-to-end localization through semantic-adaptive query generation, which reduces error propagation and improves localization precision. Compared with more recent methods, including 3D-STMN, RefMask3D, MCLN, MDIN, and IPDN, LaSA-Net still achieves the best results. The superiority of LaSA-Net mainly comes from its strong adaptability to different failure modes in complex scenes. For instance,

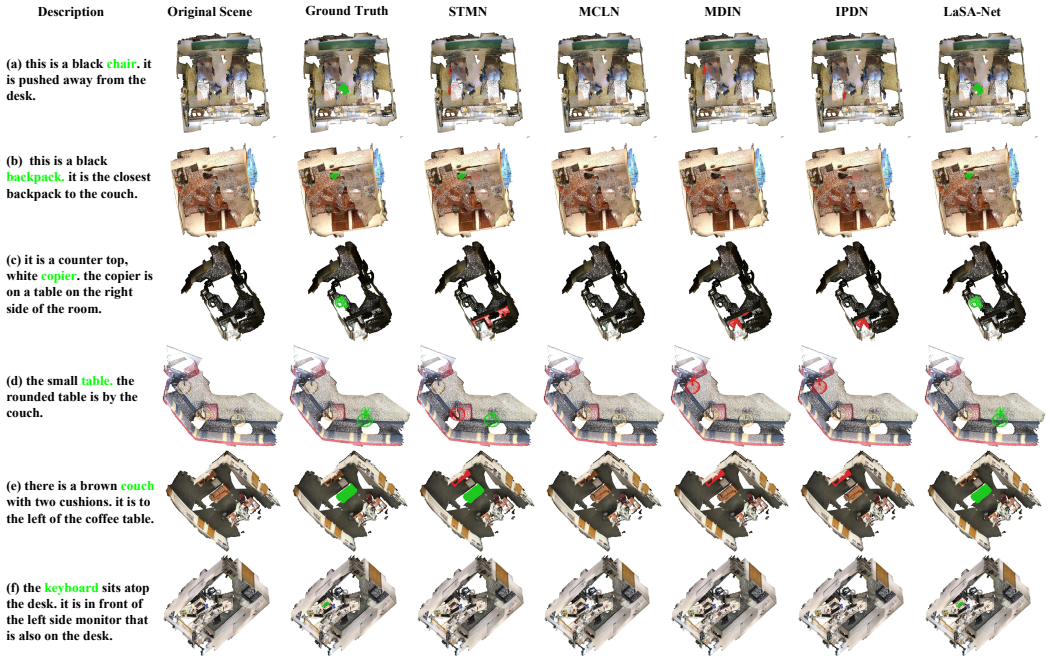

Figure 8: Qualitative comparison between SOTA method and LaSA-Net on the Scanrefer Dataset.

3D-STMN and RefMask3D are still limited by coarse hyperpoint-based object delineation, which weakens boundary discrimination between adjacent instances. MCLN may introduce inconsistency between box-level localization and mask prediction. MDIN is more vulnerable to feature dispersion when transferred from generalized multi-object grounding to single-object reference settings. And IPDN, although enhanced by image prompts, still suffers from insufficient 3D-aware alignment. In contrast, LaSA-Net jointly leverages precise 2D–3D fusion and multi-granularity sampling to preserve both fine local details and global semantic consistency, thereby yielding more accurate and robust segmentation across diverse reference scenarios.

Fig. 8 presents qualitative 3D vision–language localization results of LaSA-Net and competing methods on the ScanRefer dataset. In challenging scenes that require resolving fine-grained spatial relations and distance constraints, such as the small keyboard in Fig. 8(f) and the relational targets in Figs. 8(c) and 8(d), previous methods often suffer from different failure modes. For instance, STMN and MDIN are more easily distracted by cluttered surroundings or similar nearby objects, MCLN may produce unstable localization under coupled spatial cues, and IPDN, although able to roughly identify the target, still tends to generate coarse masks when precise boundary delineation is required. By contrast, LaSA-Net yields more accurate and complete predictions because its semantic-adaptive query generation steers query initialization toward text-relevant regions under complex relational constraints, while its multi-granularity sampling further balances local phrase alignment and global semantic consistency for targets involving cues such as direction, distance, and relative position. In scenes that additionally require joint reasoning over visual attributes and spatial topology, such as Figs. 8(a), 8(b), and 8(e), competing methods frequently localize objects at approximately correct positions but confuse critical modifiers such as color, shape, or category, revealing the limited attribute discrimination of methods relying primarily on raw 3D geometry. In contrast, LaSA-Net achieves higher segmentation fidelity by introducing dense 2D priors from the pre-trained DINOv3 ViT-H/16+ model and integrating them into precise 2D–3D fusion through geometry-aware projection and depth-consistent aggregation, thereby strengthening fine-grained cross-modal discrimination for attribute-sensitive targets in cluttered indoor scenes. Overall, these qualitative results demonstrate that LaSA-Net provides stronger robustness and precision than previous methods across diverse challenging cases, especially for small objects, relation-intensive expressions, and attribute-critical disambiguation.

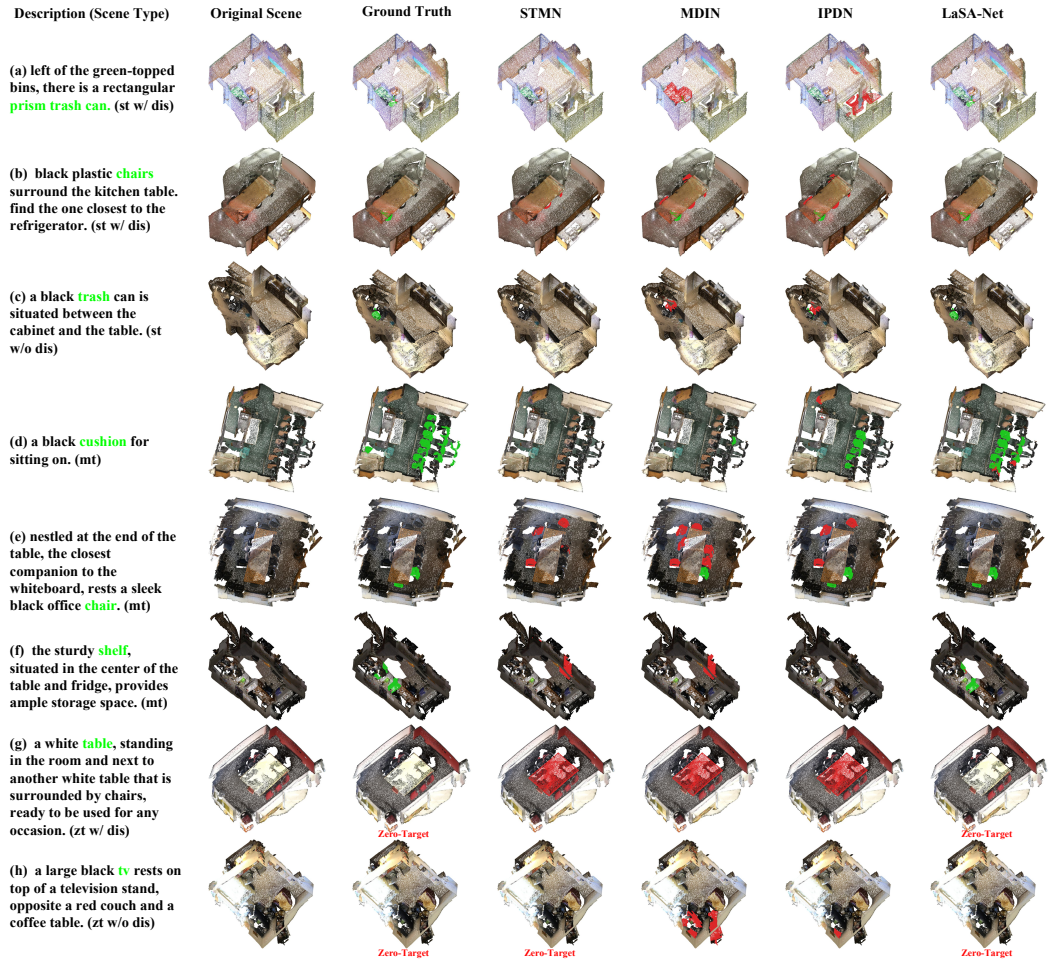

Figure 9: Qualitative comparison between SOTA method and LaSA-Net on the Multi3DRefer Dataset. ZT, ST, and MT denote zero-target, single-target, and multi-target settings, respectively. "w/ dis" and "w/o dis" indicate the cases with and without distractor objects, respectively.

F.2 Multi3DRefer Dataset

Fig. 9 compares qualitative 3D visual localization results of LaSA-Net and competitive SOTA methods on the Multi3DRefer dataset. In single-target settings, especially the distractor-heavy cases in Figs. 9(a) and 9(b), existing methods often confuse target objects with nearby visually similar instances.For example, STMN shows limited discrimination due to its direct hyperpoint–text matching paradigm, while MDIN and IPDN are more easily distracted by overlapping appearances and cluttered local contexts. By contrast, LaSA-Net more accurately isolates the referred instance, benefiting from dense 2D priors introduced by the pre-trained DINOv3 ViT-H/16+ model and their precise fusion with 3D geometry through projection-based alignment, multi-view aggregation, and depth-consistency filtering, which together enhance fine-grained appearance perception. In Fig. 9(c), where successful prediction mainly depends on correctly resolving the relational cue “between” previous methods still produce spatially shifted or incomplete masks, whereas LaSA-Net yields more accurate boundaries because its semantic-adaptive query generation explicitly biases query initialization toward text-relevant regions and reduces interference from structurally similar background points under relational constraints. In multi-target settings, as shown in Figs. 9(d)–9(f), the main challenge shifts from local discrimination to global instance coverage. STMN, MDIN, and IPDN all suffer from missed detections when valid targets are numerous and spatially dispersed, while LaSA-Net achieves more complete predictions, such as successfully localizing all 15 valid chair instances in Fig. 9(d). This advantage mainly comes from the diversity-aware seed selection strategy, which encourages

spatially distributed yet semantically relevant query initialization and thus improves coverage over discrete target instances. In zero-target settings, as shown in Figs. 9(g) and 9(h), previous methods tend to hallucinate spurious masks under invalid descriptions, revealing insufficient robustness to misleading language cues, whereas LaSA-Net consistently predicts clean empty masks. This behavior is enabled by the reliability-aware decoder, which evaluates the confidence of language guidance through cross-modal uncertainty and suppresses noisy prompt propagation when no valid referent exists. Overall, these qualitative results demonstrate that LaSA-Net achieves stronger robustness than comparative methods across diverse and challenging scenarios, including distractor interference, relational reasoning, dispersed multi-target retrieval, and hallucination-prone zero-target cases.

G Limitations

Although the proposed framework demonstrates strong performance on UrbanRefer, Multi3DRefer, and ScanRefer, we acknowledge two limitations of this work:

- Firstly, the current framework relies on high-quality cross-modal priors and accurate 2D–3D alignment, and its performance may degrade when pseudo-image quality, viewpoint coverage, or geometric calibration are poor.
- Secondly, both UrbanRefer and LaSA-Net are mainly designed for structured urban road scenes with stable geospatial cues, and their performance may be affected in rural, unstructured, or weather-degraded road environments.

In future work, we will focus on overcoming these two limitations.

H Broader Impact

In this work, we introduce LaSA-Net and UrbanRefer for outdoor generalized 3D referring expression segmentation, achieving SOTA performance on indoor and outdoor benchmarks. Looking ahead, LaSA-Net holds great promise for advancing language-guided 3D perception, benefiting downstream tasks such as autonomous driving, robotics, and urban digital twins. However, if not properly trained, point cloud encoders may learn biases from the data, reinforcing societal stereotypes. Researchers should be aware of such potential negative social impacts and continuously monitor the training data for bias.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see the Abstract and Introduction sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see the Limitations subsection in the Appendix.G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see the Method section and Experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released the dataset and code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see the Experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments are conducted under a fixed seed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see the Evaluation Criteria and Implementation Details section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see the Appendix section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see the Experiment and Reference sections.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new outdoor 3D-GRES dataset and provide code with a README file that documents the dataset link, access instructions, dataset statistics, task definitions, evaluation protocols, and usage notes.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: As described in the LLM-assisted description refinement section, an LLM was used as part of the dataset construction pipeline to refine template-generated expressions and improve linguistic diversity. The LLM was not used to generate experimental results or to replace human verification.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.