SEMANTIC AND SPATIAL DISAMBIGUATION FOR MULTI-VIEW 3D REFERRING EXPRESSION SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view 3D referring expression segmentation (MV-3DRES) aims to reconstruct a 3D scene from sparse multi-view RGB images and segment the object specified by a natural language description. However, sparse observations result in incomplete and fragmented object representations, making it difficult for existing methods to capture category-specific semantics and distinguish object instances based on their spatial locations. To address these challenges, we propose **MV-S^{2}D**, a framework that jointly mitigates semantic and spatial ambiguities in MV-3DRES. Specifically, we first introduce a Semantic Category Enhancement (SCE) strategy that strengthens category-specific representations of the target object through text-guided semantic alignment and enhancement, thereby improving semantic discrimination. To mitigate spatial instance ambiguity, we then develop a Spatial Position Guidance (SPG) mechanism that estimates text-conditioned confidence scores over scene locations to identify target-relevant regions. Finally, we design a multi-view aggregation strategy to improve the cross-view consistency of segmentation predictions. Experiments on MVRefer demonstrate that MV-S^{2}D achieves state-of-the-art performance, outperforming MVGGT by 7.3 points in Acc@0.50 on the challenging Multiple subset.

1 INTRODUCTION

3D referring expression segmentation (3D-RES) aims to locate and segment a target instance in a 3D scene given a free-form natural language description (Huang et al., 2021; Wu et al., 2024b). Unlike 3D semantic and instance segmentation, 3D-RES requires associating the category, attributes, and spatial relations expressed in language with visual evidence in the scene to distinguish the intended target from other candidates (Qian et al., 2024; Wu et al., 2024a; Wang et al., 2026a). Most existing methods (Wu et al., 2024b; He & Ding, 2024; Liu et al., 2024b) rely on dense point clouds acquired through RGB-D scanning, which requires specialized sensors and costly data collection. In contrast, sparse multi-view RGB images are easier to acquire and more practical for real-world applications. MVGGT (Wu et al., 2026b) first introduced the multi-view 3D referring expression segmentation (MV-3DRES) task, which performs language-guided 3D segmentation using sparse multi-view RGB images. Under sparse observations, the target is often represented by incomplete and fragmented visual evidence, making it difficult to capture category-specific semantics and distinguish spatially similar instances.

Sparse multi-view observations may introduce two main forms of referential ambiguity, including *semantic category ambiguity* and *spatial instance ambiguity*, as illustrated in Fig. 1. First, limited visual evidence weakens the correspondence between target-category semantics in language and the associated visual regions, leading to semantic category ambiguity (Wu et al., 2026a). Second, distinguishing instances of the same category relies on spatial cues that are often incomplete under sparse observations, resulting in spatial instance ambiguity (Wu et al., 2024a; Wang et al., 2026a). While mitigating ambiguity under sparse observations is crucial for accurate 3D referring segmentation, existing approaches still struggle with semantic category ambiguity and spatial instance ambiguity.

Based on the above observations, we propose **MV-S^{2}D**, a semantic and spatial referential disambiguation framework that alleviates ambiguity from the perspectives of semantic category and spatial

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

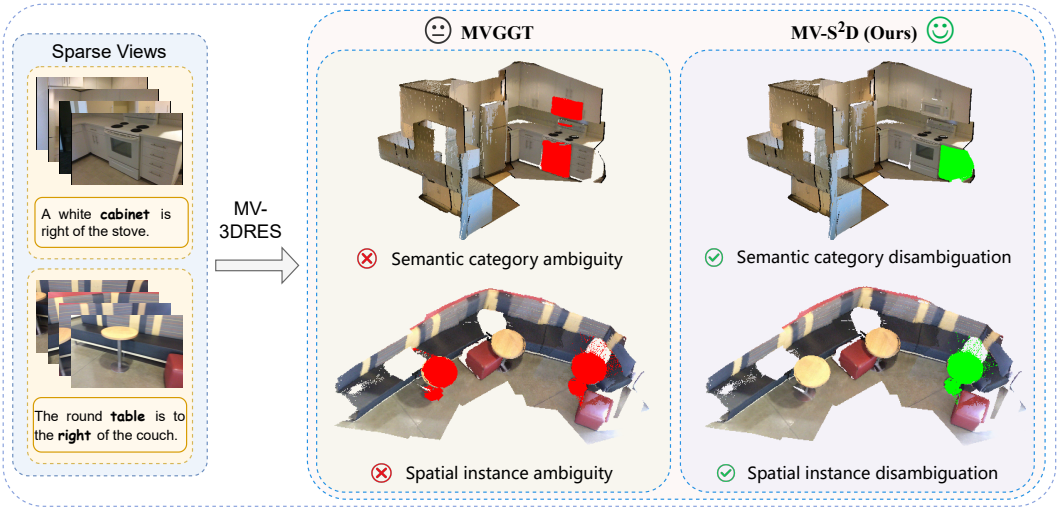


Figure 1: Comparison between our method and MVGGT. Under sparse input views, MVGGT may misidentify the target category or confuse same-category candidate instances, whereas our method leverages semantic category and spatial position cues to more accurately identify the referred target.

position. To address semantic category ambiguity, we introduce Semantic Category Enhancement (SCE), which strengthens target-category information in language through text-guided semantic category alignment and enhancement, thereby suppressing cross-category interference. To address spatial instance ambiguity, we introduce Spatial Position Guidance (SPG), which uses text to reversely filter key visual regions and leverages confidence to highlight reliable target evidence, guiding the multimodal branch to focus on regions more likely to contain the target. In addition, we design Cross-view Query Aggregation (CQA), which aggregates target information from different views through shared queries to improve cross-view consistency.

We conduct a systematic evaluation on the MVRefer benchmark. Experimental results show that MV-S^{2}D achieves state-of-the-art performance on the MV-3DRES task and obtains more pronounced improvements in challenging settings such as Multiple, validating the effectiveness of the proposed method.

Our main contributions are as follows:

- We propose MV-S^{2}D, a framework for MV-3DRES that alleviates semantic category ambiguity and spatial instance ambiguity under sparse multi-view observations.
- We introduce Semantic Category Enhancement (SCE), which selectively injects context from learnable category anchors into language representations to strengthen target-category information and reduce cross-category interference.
- We introduce Spatial Position Guidance (SPG), which combines text-to-visual relevance with intermediate mask confidence to focus visual feature updates on target-relevant regions.
- Experiments on MVRefer demonstrate that MV-S^{2}D achieves state-of-the-art performance, with particularly pronounced gains on the challenging Multiple subset.

2 Related Work

3D Referring Expression Segmentation. 3D language grounding aims to localize objects in 3D scenes according to natural language descriptions, with ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020) establishing representative benchmarks for language-guided 3D understanding. Building on this setting, 3D referring expression segmentation (3D-RES) further requires precise segmentation masks for referred targets. TGNN (Huang et al., 2021) pioneers referring 3D instance segmentation with a two-stage proposal-matching paradigm, while subsequent

methods improve candidate matching through structured cross-modal reasoning (Qian et al., 2024; Kamath et al., 2021; Liu et al., 2024a). To reduce the dependence on proposal quality, later studies increasingly adopt end-to-end segmentation based on superpoint–text matching, cross-modal interaction, spatial modeling, and query decoding (Wu et al., 2024b; He & Ding, 2024; Wu et al., 2024a). Further advances explore label-efficient learning and image- or query-enhanced decoding (Liu et al., 2024b; Chen et al., 2025). Recent works also extend the task toward fine-grained phrase-level segmentation and partial-observation settings (Chen et al., 2026; Ye et al., 2026). Nevertheless, these approaches are largely built upon pre-existing dense 3D inputs. MVGGT (Wu et al., 2026b) first introduces MV-3DRES, extending 3D-RES to sparse multi-view RGB observations.

Referential Disambiguation. Identifying the unique referent among similar candidates is a fundamental challenge in 3D referring. Existing methods improve target discrimination through semantic representation, spatial relations, and contextual reasoning. Semantic approaches enhance target representations with 2D semantics, vision–language alignment, and cross-modal modeling (Yang et al., 2021; Wu et al., 2023; Zhu et al., 2023; Wang et al., 2025; Radford et al., 2021; Kim et al., 2021; Li et al., 2021; 2022), while spatial methods distinguish similar instances through context, entity relations, and language-conditioned spatial reasoning (Yuan et al., 2021; He et al., 2021; Chen et al., 2022; Zhao et al., 2021; Luo et al., 2022; Jain et al., 2022). View-aware reasoning further exploits complementary information across views (Huang et al., 2022; Guo et al., 2023; Li et al., 2025). Recent methods increasingly combine semantic and spatial cues for fine-grained disambiguation, including category-aware and position-aware reasoning (Zhu et al., 2025; Wu et al., 2026a; Huang & Wang, 2026; Wang et al., 2026a; Tan et al., 2026). However, these approaches are mainly developed for dense 3D scenes. Under sparse multi-view observations, limited visual evidence weakens both category semantics and spatial cues, leaving semantic category ambiguity and spatial instance ambiguity insufficiently explored.

3 METHOD

3.1 PROBLEM DEFINITION

Given N sparse RGB views $\mathcal{I} = \{I_n\}_{n=1}^{N}$ and a natural language referring expression $\mathcal{T}$, the goal of MV-3DRES is to locate the instance referred to by the language from limited observations and predict the target mask corresponding to each input view. Combined with the estimated depth and camera geometry, the multi-view predictions can be further projected onto the reconstructed point cloud to obtain the 3D segmentation result of the target instance.

3.2 SEMANTIC CATEGORY ENHANCEMENT

Under sparse observations, although the target category can usually be obtained from language, limited visual information makes it difficult to establish a stable correspondence between language categories and visual regions, readily causing semantic category ambiguity. If the original language tokens are directly used to interact with visual features, the model can be affected by background regions and cross-category candidate instances. To this end, SCE explicitly introduces semantic category anchors before cross-modal interaction and strengthens target-category information through text-guided category-semantic alignment and enhancement.

Category anchor selection. We construct K learnable semantic category anchors $\mathbf{C} \in \mathbb{R}^{K\times D}$, where K denotes the number of category anchors and D is the feature dimension. In this work, we set $K = 19$. We first perform masked averaging over valid language tokens to obtain the sentence-level representation $\bar{\mathbf{e}}$, and compute its matching distribution over the category anchors:

$$\boldsymbol{\pi} = \mathrm{softmax}\big((W_q\bar{\mathbf{e}})\mathbf{C}^\top\big)\,. \tag{1}$$

where $\boldsymbol{\pi} \in \mathbb{R}^K$ denotes the matching probabilities over the K category anchors, and W_q is a learnable language projection. This soft distribution preserves uncertainty in category prediction. Category prediction is additionally supervised by the target category, while objects that cannot be mapped to predefined categories are assigned to the unknown category. The category anchors are then aggregated according to the matching distribution:

$$\mathbf{c} = W_c(\boldsymbol{\pi}\mathbf{C}). \tag{2}$$

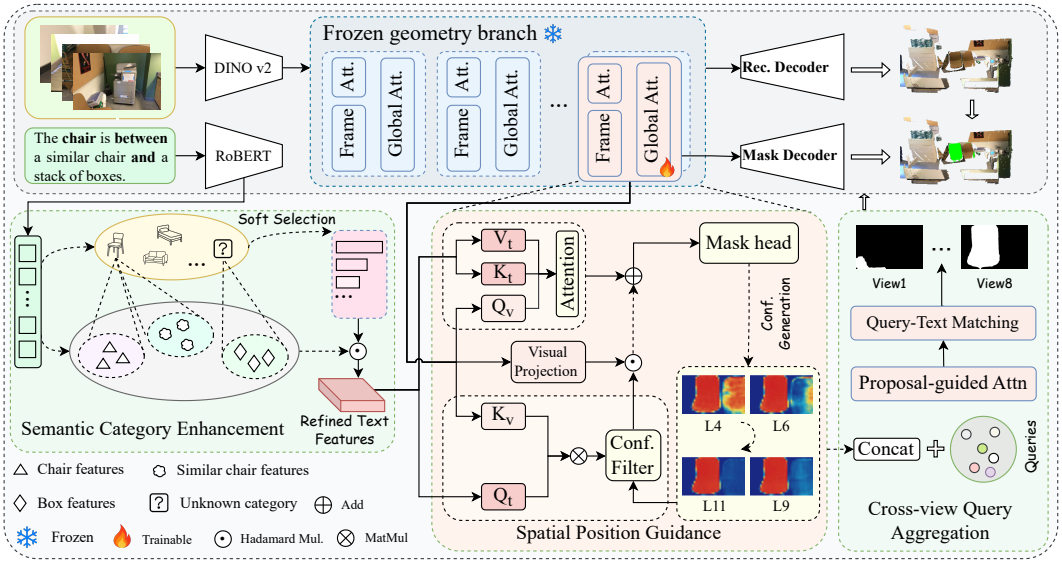


Figure 2: Overview of the proposed MV-S^{2}D framework. Given sparse multi-view RGB images and a natural language referring expression, the frozen π^3 visual–geometry branch models scene geometry and produces the reconstruction output, while the multimodal branch is initialized from its last 12 geometry-decoder layers. Semantic Category Enhancement (SCE) performs soft selection over semantic category anchors and refines the text representation. Spatial Position Guidance (SPG) combines reverse text-to-visual attention with intermediate mask confidence to generate spatial guidance and inject it into visual features. Cross-view Query Aggregation (CQA) aggregates multi-view information with shared queries.

The resulting $\mathbf{c} \in \mathbb{R}^D$ represents the category context of the current expression, and W_c denotes a learnable category projection.

Token–channel gated enhancement. Different language tokens depend on category information to different degrees. For the t-th language token $\mathbf{e}_t$, we predict a token–channel-level gating vector $\mathbf{g}_t = \sigma(\mathrm{MLP}([W_e\mathbf{e}_t; \mathbf{c}]))$, where W_e is a learnable language projection and $[\cdot;\cdot]$ denotes feature concatenation. The language representation is then updated as

$$\widetilde{\mathbf{e}}_t = \mathrm{LN}(\mathbf{e}_t + \mathbf{g}_t \odot \mathbf{c})\,. \tag{3}$$

The refined token $\widetilde{\mathbf{e}}_t$ denotes the enhanced representation of the t-th language token, and $\odot$ denotes element-wise multiplication. All refined tokens form $\widetilde{\mathbf{E}}$ for subsequent multimodal interaction.

3.3 Spatial Position Guidance

SCE can alleviate semantic category ambiguity, but when multiple candidate instances appear at different locations in the scene, category information alone may be insufficient to determine the unique target. Distinguishing these candidates further depends on visual spatial cues, which are often incomplete under sparse views. Therefore, SPG uses language information to filter visual positions and combines it with historical mask confidence to guide the multimodal branch toward the target instance.

Text-guided spatial relevance. Conventional cross-modal interaction uses visual tokens as queries and language tokens as keys and values to inject language-conditioned information into visual representations. SPG further reverses the attention direction, using language tokens as queries and visual tokens as keys to estimate the relevance of each visual position:

$$A^h_{ts} = a_t\ \mathrm{softmax}_s\Big((W^h_Q\,\mathrm{LN}(\widetilde{\mathbf{e}}_t))^\top (W^h_K\,\mathrm{LN}(\mathbf{x}_s))/(\sqrt{d_h}\,\tau)\Big)\,. \tag{4}$$

where t and s index language tokens and visual positions, respectively, and h indexes attention heads. $\mathbf{x}_s$ denotes the visual feature at position s; W_Q^h and W_K^h are the query and key projections of the h-th head, d_h is the per-head feature dimension, τ is a learnable temperature, and a_t is the valid-token mask. The resulting A_{ts}^h measures the relevance of language token t to visual position s.

We use 16 attention heads and average the responses over all valid language tokens and attention heads to obtain the language relevance r_s for each visual position.

Confident position filtering. SPG is activated after the model has formed an initial localization. Let q_s denote the mask confidence at visual position s predicted by the preceding multimodal layer. We use it as a historical localization signal through stop-gradient. Given a confidence threshold θ_{spg}, the candidate set is defined as

$$\mathcal{C} = \{s \mid q_s \geq \theta_{\text{spg}}\}. \tag{5}$$

The set $\mathcal{C}$ contains the high-confidence candidate positions, and the confidence threshold θ_{spg} is set to 0.70 in our experiments. Register tokens are excluded from $\mathcal{C}$.

The language relevance is then normalized within the candidate set:

$$w_s = \begin{cases} r_s / \sum_{j \in \mathcal{C}} r_j, & s \in \mathcal{C}, \\ 0, & s \notin \mathcal{C}. \end{cases} \tag{6}$$

The weight w_s denotes the spatial guidance weight at position s, while j indexes positions in $\mathcal{C}$.

The spatial position guidance is constructed as

$$\mathbf{P}_s = \gamma w_s \, \phi(\mathrm{LN}(\mathbf{x}_s)) \, . \tag{7}$$

The resulting $\mathbf{P}_s$ represents the spatial guidance at position s, where γ is a learnable scale and $\phi(\cdot)$ denotes the visual feature projection.

Let $\mathbf{P}$ collect the guidance features over all visual positions. The guidance is fused with the multimodal response, and the visual representation is updated through a gated residual connection:

$$\mathbf{X}^+ = \mathbf{X} + \sigma(G(\mathbf{R})) \odot \mathbf{R}, \qquad \mathbf{R} = \rho(\mathbf{F}_{\text{mm}} + \mathbf{P}) \, . \tag{8}$$

In Eq. 8, $\mathbf{X}$ and $\mathbf{X}^+$ denote the visual representations before and after the update, respectively. $\mathbf{F}_{\text{mm}}$ is the multimodal response and $\mathbf{R}$ is the fused residual feature, while $\rho(\cdot)$ and $G(\cdot)$ denote the fusion mapping and gating function, respectively.

SPG is activated from the fourth multimodal fusion layer. As decoding proceeds, the candidate regions gradually shift from broader relevant regions toward spatial positions where the target is more likely to be present.

3.4 Cross-view Query Aggregation

Different views may provide inconsistent target information. To address this, CQA uses shared queries to aggregate target information across views. After removing register tokens, patch features from all views are concatenated as $\mathbf{F} \in \mathbb{R}^{NP \times D}$. We use $K_q = 20$ shared queries, updated through two rounds of cross-attention, self-attention, and feed-forward networks:

$$\mathbf{Q}^{\ell+1} = \mathrm{FFN}\big(\mathrm{SA}\big(\mathrm{CA}(\mathbf{Q}^{\ell}, \mathbf{F})\big)\big) \, , \tag{9}$$

Here, $\mathbf{Q}^{\ell}$ represents the shared queries at the ℓ-th layer. The SCE-enhanced language representation then weights these queries to emphasize those more relevant to the current referring expression:

$$\alpha_k = \mathrm{softmax}_k\Big((W_u \mathbf{u}_k)^\top W_s \mathbf{e}_{\text{sent}} / \sqrt{D}\Big) \, , \tag{10}$$

where α_k denotes the relevance of the k-th query to the expression. Finally, the query responses are weighted and upsampled to obtain $\mathbf{M}_{\text{CQA}}$, which is residually fused with the original prediction using a learnable gate. This exploits complementary information across views to improve prediction consistency.

3.5 Training Objective

During training, the visual–geometry reconstruction branch is frozen and used only to provide 3D structural information. The training objective includes final segmentation, deep supervision, semantic category supervision, and cross-view aggregation supervision. The final mask is optimized with binary cross-entropy and per-view Dice losses, with reduced weights for target-absent views, while intermediate predictions and SCE use deep and category supervision, respectively. For CQA, auxiliary segmentation, background suppression, and patch-level contrastive losses are jointly denoted as $\mathcal{L}_{\mathrm{CQA}}$. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{deep}}\mathcal{L}_{\mathrm{deep}} + \lambda_{\mathrm{sce}}\mathcal{L}_{\mathrm{SCE}} + \lambda_{\mathrm{cqa}}\mathcal{L}_{\mathrm{CQA}}. \tag{11}$$

Here, $\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{deep}}$, $\mathcal{L}_{\mathrm{SCE}}$, and $\mathcal{L}_{\mathrm{CQA}}$ denote the final segmentation, deep supervision, semantic category supervision, and cross-view aggregation losses, respectively. The main paper retains ablations on the core components, SPG candidate confidence threshold, SPG starting fusion layer, SCE category-supervision weight, and the number of shared CQA queries.

4 Experiments

4.1 Experimental Setup

Dataset and evaluation metrics. We evaluate our method on the MVRefer benchmark, built from ScanRefer annotations (Chen et al., 2020) and ScanNet RGB sequences (Dai et al., 2017). Each sample contains 8 sparse RGB views with at least one target-visible view. Following MVGGT (Wu et al., 2026b), we use the same dataset split and evaluation protocol for fair comparison. We report Global, View, Pos, and Neg mIoU, together with Acc@0.25 and Acc@0.50, and evaluate performance on the Hard/Easy and Unique/Multiple subsets.

Implementation details. We use pretrained π^3 (Wang et al., 2026b) as the visual–geometry backbone and initialize the multimodal branch from its last 12 decoder layers, with a feature dimension of 1024. Text is encoded by RoBERTa-base, while the visual–geometry branch is frozen during training. We randomly sample 4 views for training and use all 8 views for testing. Main experiments use a training resolution of 448×448, ablations use 224×224, and all methods are evaluated at 518×336.

4.2 Main Results

Overall, MV-S^{2}D achieves consistent improvements on MVRefer, reaching **43.2** Overall Global mIoU and **72.7** View mIoU, outperforming MVGGT (Wu et al., 2026b) by 3.3 and 3.4 percentage points, respectively. Notably, Global mIoU improves by 4.8 points on Hard and 3.6 points on Multiple, indicating that semantic-category and spatial-position disambiguation is particularly effective when target information is incomplete or multiple same-category candidates are present.

Comparison on the MVRefer benchmark. Table 1 compares different methods on the Hard, Easy, Unique, Multiple, and Overall subsets, reporting Global, View, Pos, and Neg mIoU following the MVRefer protocol.

Comparison under different IoU thresholds. We further report Acc@0.25, Acc@0.50, and mIoU following the ScanRefer protocol (Chen et al., 2020). Since traditional 3D-RES uses ground-truth point clouds while MV-3DRES uses only sparse RGB images, Table 2 lists the two settings separately.

Under the stricter IoU threshold, MV-S^{2}D shows a more pronounced advantage. Overall Acc@0.50 increases from 41.5 to **47.8**, with a larger gain than Acc@0.25, indicating more accurate target masks. On the challenging Multiple subset, Acc@0.50 improves by 7.3 percentage points, demonstrating the effectiveness of joint semantic and spatial disambiguation in resolving multiple candidate instances. Acc@0.50 on Unique also increases from 74.5 to **77.0** (+2.5 points), showing larger benefits in ambiguous scenes while maintaining consistent gains in simpler cases. Notably, with only sparse RGB inputs, MV-S^{2}D surpasses several traditional 3D-RES methods using ground-truth

Table 1: Performance comparison on the MVRefer benchmark. All metrics are mIoU under different categories. Values are reported as percentages (%), and the best results are shown in bold.

Method	Hard (∼ 40%)				Easy (∼ 60%)				Unique (∼ 19%)				Multiple (∼ 81%)				Overall			
	Global	View	Pos	Neg	Global	View	Pos	Neg	Global	View	Pos	Neg	Global	View	Pos	Neg	Global	View	Pos	Neg
Two-stage (Wu et al., 2026b)	8.1	8.6	22.6	10.0	25.8	28.2	45.6	21.6	27.5	41.1	52.1	30.1	16.4	15.4	32.1	13.8	18.5	20.3	35.9	16.9
2D-Lift (Wu et al., 2026b)	6.4	15.0	25.3	11.9	25.4	24.1	45.1	12.3	31.5	30.9	47.2	19.2	14.5	17.9	34.8	10.4	17.8	20.4	37.2	12.1
MVGGT (Wu et al., 2026b)	24.4	67.3	31.6	78.0	50.1	70.6	**52.2**	81.2	65.2	82.6	64.5	90.4	33.8	66.1	39.0	77.4	39.9	69.3	44.0	79.9
MV-S^{2}D (Ours)	**29.2**	**72.0**	**33.8**	**83.4**	**52.5**	**73.1**	52.1	**85.0**	**67.1**	**84.1**	**64.7**	**92.5**	**37.4**	**69.9**	**40.0**	**82.4**	**43.2**	**72.7**	**44.8**	**84.4**

Table 2: Performance comparison under the original ScanRefer protocol. Traditional 3D-RES methods use ground-truth point clouds, whereas MV-3DRES methods use only sparse RGB inputs. All values are percentages (%).

Method	Unique (∼ 19%)			Multiple (∼ 81%)			Overall		
	Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU
Traditional 3D-RES (with ground-truth point clouds)									
TGNN (Huang et al., 2021)	69.3	57.8	50.7	31.2	26.6	23.6	38.6	32.7	28.8
3D-STMN (Wu et al., 2024b)	89.3	84.0	74.5	46.2	29.2	31.1	54.6	39.8	39.5
SegPoint (He et al., 2024)	–	–	–	–	–	–	–	–	41.7
Reason3D (Huang et al., 2025)	88.4	84.2	74.6	50.5	31.7	34.1	57.9	41.9	42.0
RG-SAN (Wu et al., 2024a)	89.2	84.3	74.5	55.0	35.4	37.4	61.7	44.9	44.6
LESS (Liu et al., 2024b)	–	–	–	–	–	–	53.2	29.9	33.7
3D-LLaVA (Deng et al., 2025)	–	–	–	–	–	–	–	–	43.3
MV-3DRES (sparse RGB inputs only)									
Two-stage (Wu et al., 2026b)	43.8	20.9	27.5	25.1	12.2	16.4	28.7	13.9	18.5
2D-Lift (Wu et al., 2026b)	51.6	26.0	31.5	21.5	5.6	14.5	27.3	9.6	17.8
MVGGT (Wu et al., 2026b)	83.6	74.5	65.2	49.2	33.5	33.8	55.9	41.5	39.9
MV-S^{2}D (Ours)	**85.4**	**77.0**	**67.1**	**51.2**	**40.8**	**37.4**	**57.9**	**47.8**	**43.2**

point clouds in Overall mIoU, demonstrating strong competitiveness under more challenging input conditions.

4.3 Qualitative Results

Figure 3 shows qualitative results on MVRefer. Each column corresponds to one language–target example, showing the reconstructed scene, ground-truth target, MVGGT prediction, and our prediction from top to bottom.

Under sparse multi-view observations, incomplete scenes and visually or semantically similar candidates can interfere with target identification. MVGGT may select an incorrect instance due to inaccurate category recognition or spatial-relation reasoning, or recover only part of the target region. In contrast, MV-S^{2}D strengthens target-category semantics and exploits spatial constraints, leading to more reliable candidate discrimination and more complete target segmentation. These results support the effectiveness of semantic and spatial disambiguation in sparse and complex scenes.

4.4 Ablation Studies

We analyze the effects of individual components and key designs on the MVRefer validation set. To reduce the computational cost of ablation studies, all experiments use the same 224×224 training resolution as MVGGT, with identical testing settings and evaluation protocols. Under this setting, our full model still outperforms MVGGT by 2.9 points in Global mIoU. All subsequent ablations are conducted under the same setting for fair comparison.

Core component analysis. Table 3 analyzes the contributions of SCE, SPG, and CQA. The full model achieves the best performance, while removing any component causes a performance drop, indicating their complementary effects. The larger gains on Hard and Multiple further show that joint modeling is particularly effective under sparse observations and multiple candidate instances.

Comparison with adapted disambiguation strategies. To examine the transferability of existing disambiguation strategies to MV-3DRES, we adapt representative designs from IPDN (Chen et al., 2025), Position3D (Wang et al., 2026a), and EG-3DVG (Park et al., 2026) to the same CQA-only baseline. Their core mechanisms are reformulated from superpoint-based 3D representations to multi-view visual tokens. As shown in Table 6, these adaptations do not outperform the CQA-only

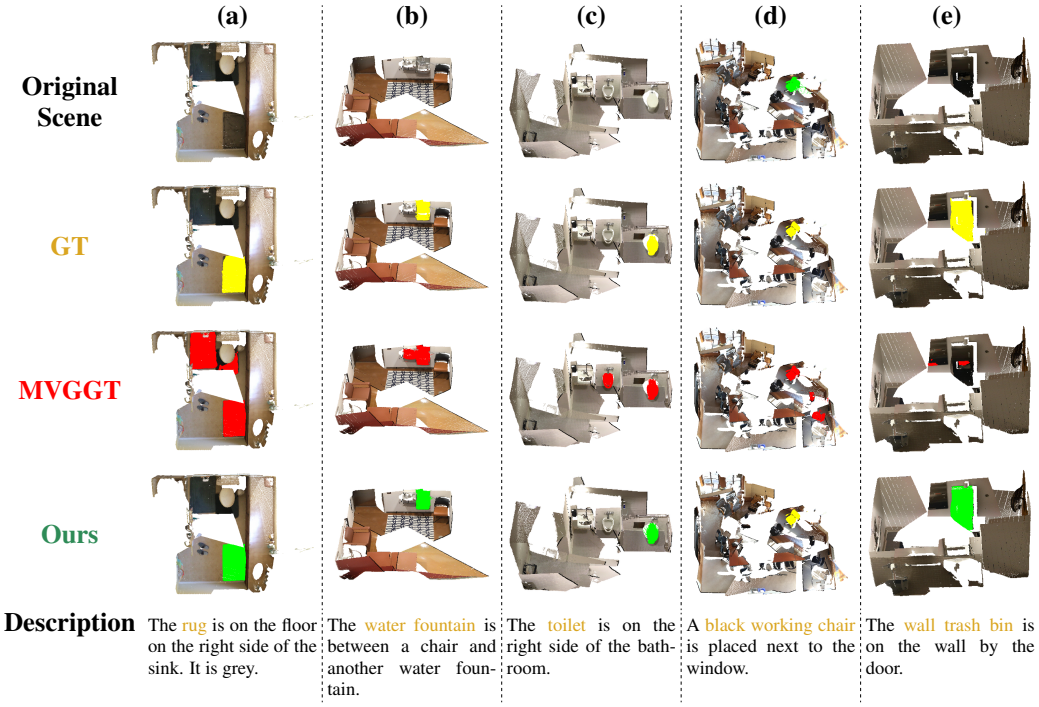


Figure 3: Qualitative comparison on MVRefer. Each column shows one referring example. From top to bottom: the reconstructed scene (Original Scene), the ground-truth target annotation projected onto the reconstructed point cloud (GT), the MVGGT prediction, and our prediction. The referring expressions are shown at the bottom, with the target noun phrase highlighted in green.

Table 3: Component ablation of SCE, SPG, and CQA with MVGGT as baseline.

Components			Global mIoU		Overall mIoU			
SCE	SPG	CQA	Hard	Multiple	Global	View	Pos	Neg
–	–	–	24.4	33.8	39.9	69.3	44.0	79.9
✓	–	–	26.1	35.6	41.5	71.2	44.4	82.4
–	✓	–	24.6	34.4	40.5	69.7	44.4	79.9
–	–	✓	26.0	35.4	41.5	70.3	44.8	80.5
–	✓	✓	25.6	35.3	41.4	69.8	45.3	79.8
✓	–	✓	26.3	35.7	41.7	70.4	45.3	80.7
✓	✓	–	26.7	35.8	41.6	70.6	45.1	81.2
✓	✓	✓	**27.2**	**36.9**	**42.8**	**72.2**	**45.8**	**82.9**

Table 4: Ablation on the candidate confidence threshold θ_{spg} in SPG.

θ_{spg}	Hard mIoU	Multiple mIoU	Acc@0.25	Acc@0.50	mIoU
0.3	26.2	35.8	58.1	44.7	41.9
0.5	26.7	36.4	58.6	45.5	42.2
0.7	**27.2**	**36.9**	**59.1**	**46.3**	**42.8**
0.9	26.7	35.9	57.8	44.6	41.7

Table 5: Ablation on the starting fusion layer in SPG.

Start Layer	Hard mIoU	Multiple mIoU	Acc@0.25	Acc@0.50	mIoU
3	27.0	36.3	58.0	45.1	42.1
4	**27.2**	**36.9**	**59.1**	**46.3**	**42.8**
5	26.7	36.7	58.5	45.4	42.5
6	27.1	36.6	58.5	45.5	42.4

baseline, indicating limited transferability to sparse multi-view scenes. In contrast, our SCE and SPG achieve better performance.

Semantic category anchor visualization. Fig. 4 further analyzes the category anchors learned by SCE. For different referring expressions, higher matching weights are concentrated on the target and semantically related categories, while irrelevant categories receive low responses. This shows that SCE can retrieve target-relevant category semantics from the learned anchors, providing more discriminative category context for subsequent multimodal interaction.

SPG design choices. As shown in Tables 4 and 5, $\theta_{\text{spg}} = 0.7$ performs best, showing that moderate filtering reduces irrelevant responses, while an overly high threshold suppresses valid ones. SPG also performs best when activated from the fourth fusion layer, whereas earlier or later activation weakens position guidance.



The white cabinets are below the counter with a silver sink.

The dark brown cabinet is under the sink and left of the dishwasher.

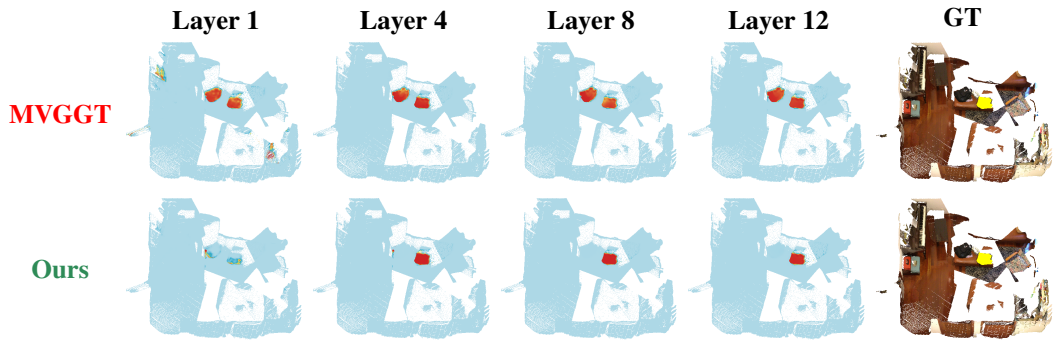
Figure 4: Visualization of semantic category anchor responses in SCE. The learned anchors exhibit category-selective responses, indicating that the learnable category anchors capture category-specific semantics.

Table 6: Comparison with representative 3D-RES disambiguation strategies adapted to MV-3DRES. All strategies are evaluated on the same CQA-only baseline.

Method	Hard	Multiple	Acc@0.25	Acc@0.50	mIoU
CQA only	26.0	35.4	57.6	44.4	41.5
+ IPDN (Chen et al., 2025)	25.7	34.8	56.8	43.9	41.1
+ EG-3DVG (Park et al., 2026)	26.0	35.3	57.7	44.1	41.4
+ Position3D (Wang et al., 2026a)	25.7	35.5	57.3	44.5	41.4
+ SCE + SPG (Ours)	**27.2**	**36.9**	**59.1**	**46.3**	**42.8**

Table 7: Acc@0.50 (%) on spatial and non-spatial subsets.

Method	Spatial	Non-spatial
Proportion	78.70%	21.30%
w/o CQA	44.0	44.7
w/o SCE	43.6	44.8
w/o SPG	44.0	46.9
MV-S^{2}D	**46.2**	**46.9**



Description: *"It is a small black backpack. It is sitting on a table between the couch and the leather chair."*

Figure 5: Layer-wise visualization of the spatial disambiguation effect of SPG. The two rows show the spatial responses of MVGGT and our method at multimodal fusion layers 1, 4, 8, and 12, respectively. The last column shows the ground-truth target. Compared with MVGGT, SPG progressively suppresses distractor regions and focuses on the target.

Spatial disambiguation analysis. Table 7 and Fig. 5 analyze SPG from quantitative and visual perspectives. Removing SPG reduces Acc@0.50 on the Spatial subset from 46.2 to 44.0, while leaving the Non-spatial subset unchanged. Meanwhile, MVGGT produces dispersed responses over multiple candidate regions, whereas our method progressively suppresses distractors and focuses on the target. These results demonstrate the effectiveness of SPG in spatial disambiguation.

5 Conclusion

We present MV-S^{2}D, a semantic and spatial disambiguation framework for 3D referring expression segmentation under sparse multi-view observations. SCE and SPG exploit category semantics and spatial cues for referential disambiguation, while CQA aggregates complementary information across views. Experiments on MVRefer demonstrate state-of-the-art performance, with larger gains on the challenging Multiple subset. These results show that jointly exploiting semantic, spatial, and cross-view information is effective for referential disambiguation under sparse observations. Future research may explore more general multi-view representation learning, adaptive view aggregation, and broader language-guided 3D understanding tasks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language translation, literature organization, and manuscript editing. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

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*, pp. 422–440. Springer, 2020.

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

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, pp. 2132–2140, 2025.

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, pp. 2966–2974, 2026.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3772–3782. IEEE, 2025.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15326–15337. IEEE, 2023.

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*, pp. 2344–2352, 2021.

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

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

Jiale Huang and Shangfei Wang. Saqn: Semantic-based adaptive query network for 3d referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38627–38636, 2026.

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)*, pp. 1177–1186. IEEE, 2025.

Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 1610–1618, 2021.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15503–15512. IEEE, 2022.

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

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 1760–1770. IEEE, 2021.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3707–3717. IEEE, 2025.

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

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, 2024b.

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)*, pp. 16433–16442. IEEE, 2022.

GwangWook Park, Hyo-Jun Lee, Jong-Hyeon Baek, Hanul Kim, and Yeong Jun Koh. Eg-3dvg: Expression and geometry aware grounding decoder for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2026.

Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4551–4559, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Wenbin Tan, Jiawen Lin, Fangyong Wang, Yuan Xie, Yong Xie, Yachao Zhang, and Yanyun Qu. Pc-crossdiff: Point-cluster dual-level cross-modal differential attention for unified 3d referring and segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9332–9340, 2026.

Yabing Wang, Zhuotao Tian, Le Wang, Zheng Qin, and Sanping Zhou. Spatial matters: Position-guided 3d referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39486–39496, 2026a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

Yuan Wang, Ya-Li Li, WU Eastman ZY, and Shengjin Wang. Liba: Language instructed multi-granularity bridge assistant for 3d visual grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 8114–8122, 2025.

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, 2024a.

Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5940–5948, 2024b.

Changli Wu, Qi Chen, Jiayi Ji, Yihang Liu, Yiwei Ma, Xiaoshuai Sun, and Liujuan Cao. 3d-stmn++: Leveraging semantic proxies to enhance superpoint-text matching for 3d referring expression segmentation. *Pattern Recognition*, pp. 113854, 2026a.

Changli Wu, Haodong Wang, Jiayi Ji, Yutian Yao, Chunsai Du, Jihua Kang, Yanwei Fu, and Liujuan Cao. Mvggt: Multimodal visual geometry grounded transformer for multiview 3d referring expression segmentation. *arXiv preprint arXiv:2601.06874*, 2026b.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19231–19242. IEEE, 2023.

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)*, pp. 1836–1846. IEEE, 2021.

Xulun Ye, Qin Zhang, and Kun Zhou. Langref3dgs: Natural language-guided 3d referential segmentation from partial observations via 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38595–38605, 2026.

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)*, pp. 1771–1780. IEEE, 2021.

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)*, pp. 2908–2917. IEEE, 2021.

Yihang Zhu, Jinhao Zhang, Yuxuan Wang, Aming Wu, and Cheng Deng. Vgmamba: Attribute-to-location clue reasoning for quantity-agnostic 3d visual grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5295–5304. IEEE, 2025.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2899–2909. IEEE, 2023.

APPENDIX

A TRAINING DETAILS

Training configuration. We train MV-S^{2}D for 30 epochs with 1,000 iterations per epoch, resulting in 30,000 optimization steps in total. During training, each referring-expression sample contains four RGB views at a resolution of 448×448. At inference time, all eight sparse views are used and resized to 518×336, following the evaluation setting of MVRefer. The visual–geometry reconstruction branch initialized from the pretrained π^3 model and the language encoder are frozen throughout training, while only the multimodal referring-segmentation modules, including SCE, SPG, and CQA, are optimized.

Batch construction and view sampling. We employ a view-aware dynamic batch sampler with a maximum budget of 64 images per GPU. Since each training sample contains exactly four views, this corresponds to 16 referring-expression samples, or 64 RGB images, in each optimization iteration. No gradient accumulation is used.

For each training sample, we explicitly balance target-visible and target-absent observations. Specifically, among the four sampled views, two views are sampled from frames in which the referred instance is visible and the remaining two are sampled from frames in which it is absent. If insufficient target-absent frames are available, target-visible frames are sampled with replacement to maintain the fixed number of input views. The selected views are randomly shuffled before being fed to the network. This sampling strategy exposes the model to both positive visual evidence and challenging negative views, encouraging it to suppress irrelevant regions when the referred object is not observable.

Data augmentation. To improve robustness to appearance variations and image degradation, we apply several image augmentations during training. Color jitter is independently applied to brightness, contrast, saturation, and gamma within $[0.7, 1.3]$, while the hue is perturbed within $[-0.1, 0.1]$. JPEG degradation is applied with probability 0.5, with the compression quality randomly sampled between 20 and 100. We additionally apply image blurring with probability 0.5 by first downsampling the image with a random scale in $[0.25, 1.0]$ and then resizing it back to the original resolution. The training pipeline further employs mild crop and focal-length perturbations. The corresponding instance masks and camera intrinsics are transformed consistently with the RGB images.

Optimization. We optimize the trainable parameters using AdamW with a peak learning rate of 1×10^{-4}, a weight decay of 5×10^{-2}, and $\beta_1 = 0.9$, $\beta_2 = 0.95$. Bias terms and one-dimensional parameters are excluded from weight decay. The learning rate is controlled by a OneCycle schedule over the entire 30,000-step training process, using cosine annealing. We set the OneCycle peak learning rate to 1×10^{-4}, with `pct_start`$= 0$, `div_factor`$= 1000$, and `final_div_factor`$= 100$.

Training objective. The final referring mask is supervised using binary cross-entropy (BCE) and Dice losses with equal weights:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}. \tag{12}$$

The Dice loss is computed independently for each view. For target-absent views, their contributions are reweighted according to the number of negative views in the corresponding sample, preventing multiple empty views from dominating the optimization.

In addition to the final prediction, intermediate mask predictions from the multimodal decoder are supervised using the same BCE–Dice objective. The deep-supervision loss is therefore written as

$$\mathcal{L}_{\text{deep}} = \sum_{\ell} \left(\mathcal{L}^{(\ell)}_{\text{BCE}} + \mathcal{L}^{(\ell)}_{\text{Dice}} \right), \tag{13}$$

where ℓ denotes an intermediate multimodal fusion layer. Each intermediate prediction is weighted by 0.5 in the total objective.

For SCE, the semantic category-anchor distribution is explicitly supervised with the target semantic category using a cross-entropy loss, denoted as $\mathcal{L}_{\mathrm{SCE}}$. Its loss weight is set to

$$\lambda_{\mathrm{SCE}} = 0.04. \tag{14}$$

CQA is trained with three auxiliary constraints. First, its aggregated mask prediction is supervised using the same BCE–Dice segmentation objective:

$$\mathcal{L}_{\mathrm{CQA}}^{\mathrm{aux}} = \mathcal{L}_{\mathrm{BCE}}^{\mathrm{CQA}} + \mathcal{L}_{\mathrm{Dice}}^{\mathrm{CQA}}. \tag{15}$$

We assign this auxiliary segmentation loss a weight of 0.3.

Second, we introduce a background-suppression loss that penalizes predicted foreground probability outside the ground-truth target:

$$\mathcal{L}_{\mathrm{bg}} = \frac{1}{|\Omega|} \sum_{s \in \Omega} \sigma(m_s)(1 - y_s), \tag{16}$$

where m_s and y_s denote the predicted logit and binary target label at position s, respectively. Its maximum weight is 0.03 and is linearly warmed up during the first five epochs:

$$\lambda_{\mathrm{bg}}(e) = 0.03 \min\left(\frac{e+1}{5}, 1\right), \tag{17}$$

where e denotes the zero-indexed training epoch.

Third, a patch-level contrastive objective $\mathcal{L}_{\mathrm{p2p}}$ encourages target patches to form a compact representation while separating them from background patches. Its weight is set to 0.02. We use a contrastive temperature of 0.05 and sample at most 1,024 positive and 1,024 negative patch features per sample.

Accordingly, the CQA-related objective is

$$\mathcal{L}_{\mathrm{CQA}} = 0.3\,\mathcal{L}_{\mathrm{CQA}}^{\mathrm{aux}} + \lambda_{\mathrm{bg}}(e)\mathcal{L}_{\mathrm{bg}} + 0.02\,\mathcal{L}_{\mathrm{p2p}}. \tag{18}$$

The complete training objective can therefore be expanded as

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + 0.5\,\mathcal{L}_{\mathrm{deep}} + 0.04\,\mathcal{L}_{\mathrm{SCE}} + \mathcal{L}_{\mathrm{CQA}}. \tag{19}$$

No superpoint-level segmentation or consistency loss is enabled in our final configuration.

Reproducibility. We use deterministic training whenever supported by the underlying operators. The global random seed is set to 1111, while the data sampler uses a base seed of 666 and is reseeded at every epoch. The training loader uses 16 workers and pinned memory. Model checkpoints and training statistics are recorded throughout training using TensorBoard. Unless otherwise specified, all main experiments follow the above configuration.

B Model Architecture Details

This section provides additional implementation details of MV-S^{2}D. We first describe the visual–geometry branch, the multimodal branch, and their feature interaction, and then provide detailed implementations of SCE, SPG, and CQA. Unless otherwise specified, we follow the notation used in the main paper.

Visual–geometry backbone and token organization. The visual encoder is implemented with DINOv2 ViT-L/14 with a patch size of 14 and feature dimension $D = 1024$. For an input image of spatial size $H \times W$, each view produces

$$P = \frac{H}{14}\frac{W}{14} \tag{20}$$

patch tokens. We further prepend $R = 5$ learnable register tokens to each view, where R denotes the number of register tokens. Accordingly, the initial visual representation of the n-th view is

$$\mathbf{V}_n^0 \in \mathbb{R}^{(P+R)\times D}, \qquad n = 1, \ldots, N, \tag{21}$$

where $\mathbf{V}_n^0$ denotes the encoded visual tokens of view n before the geometry decoder.

The geometry decoder contains 36 Transformer blocks with 16 attention heads, giving a per-head dimension of $d_h = 64$. Two-dimensional rotary positional encoding is applied to spatial patch tokens, while register tokens do not receive spatial coordinates. The decoder alternates between intra-view and cross-view modeling. Specifically, using zero-based decoder indices, an even-indexed block processes each view independently, whereas an odd-indexed block jointly processes the concatenated tokens from all N views. This alternating organization preserves view-specific structure while propagating information across different views.

Language representation. The referring expression $\mathcal{T}$ is encoded using RoBERTa-base. Its 768-dimensional token features are projected to the common $D = 1024$ feature space through a learnable linear projection. For an expression containing L language tokens, we write

$$\mathbf{E} = \{\mathbf{e}_t\}_{t=1}^{L} = W_{\text{text}}\,\text{RoBERTa}(\mathcal{T}), \qquad \mathbf{E} \in \mathbb{R}^{L\times D}, \tag{22}$$

where L denotes the number of language tokens and W_{text} denotes the projection from the RoBERTa feature space to the multimodal feature space. The notation $\mathbf{e}_t$ is consistent with the language token used in SCE in the main paper.

Multimodal decoder. The multimodal decoder consists of 12 Transformer blocks initialized from the last 12 layers of the pretrained π^3 geometry decoder. It therefore retains the same $D = 1024$ feature dimension, 16-head attention configuration, and alternating intra-view/cross-view organization as the corresponding geometry layers. During referring-segmentation training, the visual–geometry branch remains frozen, while the multimodal decoder is optimized.

At each multimodal layer, the primary cross-modal interaction uses visual tokens as queries and the SCE-enhanced language representation $\widetilde{\mathbf{E}}$ as keys and values. In contrast, SPG reverses the attention direction and uses language tokens as queries and visual tokens as keys to estimate spatial relevance. The reverse attention does not replace the main visual-query/text-key-value fusion stream, but instead provides an additional spatial guidance signal.

Let $\mathbf{G}^\ell$ denote the frozen geometry feature associated with the ℓ-th multimodal layer, where $\ell = 1, \ldots, 12$. We use $\mathbf{X}^{+,\ell}$ to denote the multimodal visual representation after the language-conditioned update at layer ℓ. Each layer first predicts an intermediate referring mask from $\mathbf{X}^{+,\ell}$ and subsequently injects the corresponding geometry feature for further decoding.

Intermediate mask prediction and geometry injection. Register tokens are removed before mask prediction. The mask head contains two 3×3 convolutional layers followed by a 1×1 output layer, with ReLU activation after the first two convolutions. Let $\mathbf{M}^\ell$ denote the intermediate mask logits predicted at the ℓ-th multimodal layer. Then

$$\mathbf{M}^\ell = \text{Up}\Big(\mathcal{H}_{\text{mask}}\left(\mathbf{X}^{+,\ell}_{\text{patch}}\right)\Big), \tag{23}$$

where $\mathbf{X}^{+,\ell}_{\text{patch}}$ denotes the patch-token subset of $\mathbf{X}^{+,\ell}$ and $\text{Up}(\cdot)$ denotes bilinear upsampling. The mask prediction head is

$$\mathcal{H}_{\text{mask}} = \text{Conv}_{1\times1} \circ \text{ReLU} \circ \text{Conv}_{3\times3} \circ \text{ReLU} \circ \text{Conv}_{3\times3}\,. \tag{24}$$

After predicting $\mathbf{M}^\ell$, the corresponding frozen geometry feature is injected through

$$\mathbf{X}^{\ell}_{\text{out}} = \mathbf{X}^{+,\ell} + Z^\ell\big(\mathbf{G}^\ell\big)\,, \tag{25}$$

where $Z^\ell : \mathbb{R}^D \to \mathbb{R}^D$ denotes a learnable linear projection whose weight and bias are initialized to zero, and $\mathbf{X}^{\ell}_{\text{out}}$ denotes the feature passed to the next multimodal layer. Zero initialization prevents geometry injection from perturbing the pretrained representation at the beginning of training while allowing the model to progressively exploit geometric residual information.

B.1 Semantic Category Enhancement

SCE follows the formulation in the main paper. We use $K = 19$ learnable semantic category anchors $\mathbf{C} \in \mathbb{R}^{K\times D}$ with $D = 1024$. The anchors are initialized from a zero-mean Gaussian distribution with standard deviation 0.02.

Following the notation of the main paper, valid language tokens are first averaged to obtain $\bar{\mathbf{e}}$, and the anchor matching distribution and category context are computed as

$$\boldsymbol{\pi} = \mathrm{softmax}\big((W_q\bar{\mathbf{e}})\mathbf{C}^\top\big), \qquad \mathbf{c} = W_c(\boldsymbol{\pi}\mathbf{C}). \tag{26}$$

The soft aggregation preserves category uncertainty and allows semantically related anchors to jointly contribute to the category context $\mathbf{c}$.

The token–channel gate in the main paper is implemented with a two-layer MLP:

$$\mathbf{g}_t = \sigma(W_2\ \mathrm{Drop}\left(\mathrm{GELU}\left(W_1[W_e\mathbf{e}_t;\mathbf{c}]\right)\right)), \tag{27}$$

where $W_1 \in \mathbb{R}^{D\times 2D}$ and $W_2 \in \mathbb{R}^{D\times D}$ denote the two projection layers of the gating MLP, and $\mathrm{Drop}(\cdot)$ denotes dropout with rate 0.1. The language token is then refined as

$$\widetilde{\mathbf{e}}_t = \mathrm{LN}(\mathbf{e}_t + \mathbf{g}_t \odot \mathbf{c}), \tag{28}$$

consistent with the formulation in the main paper. All refined tokens form $\widetilde{\mathbf{E}}$ and are subsequently used by the multimodal decoder, SPG, and CQA.

B.2 Spatial Position Guidance

SPG operates in parallel with the main visual-query/text-key-value cross-modal interaction. The reverse attention uses SCE-enhanced language tokens as queries and visual tokens as keys. We use 16 attention heads, giving $d_h = D/16 = 64$. The temperature τ is learnable and initialized to 1.0.

The historical mask confidence q_s is obtained from the intermediate mask predicted by the preceding multimodal layer:

$$q_s^{\ell-1} = \mathrm{sg}\big[\sigma\big(\mathrm{Down}(\mathbf{M}^{\ell-1})_s\big)\big], \tag{29}$$

where $\mathrm{Down}(\cdot)$ denotes bilinear downsampling from the image resolution to the patch grid and $\mathrm{sg}[\cdot]$ denotes stop-gradient. Register tokens are assigned zero confidence and therefore do not participate in spatial candidate selection.

The confident candidate set is defined as

$$\mathcal{C} = \{s \mid q_s \geq \theta_{\mathrm{spg}}\}, \qquad \theta_{\mathrm{spg}} = 0.70. \tag{30}$$

SPG is activated from the fourth multimodal fusion layer, allowing the first three layers to establish an initial localization. If no spatial position satisfies the confidence threshold, the SPG branch produces zero spatial guidance instead of forcing an unreliable candidate.

The visual projection $\phi(\cdot)$ is implemented with two bias-free linear layers and a GELU activation. After computing the text-conditioned relevance r_s and normalized candidate weight w_s, the spatial guidance is constructed as

$$\mathbf{P}_s = \gamma w_s \phi(\mathrm{LN}(\mathbf{x}_s)), \tag{31}$$

where the learnable residual scale γ is initialized to 0.05. The resulting $\mathbf{P}$ is combined with the original multimodal response $\mathbf{F}_{\mathrm{mm}}$ and updates the visual representation using the gated residual formulation in the main paper.

Accordingly, an SPG-enabled multimodal layer sequentially performs the multimodal Transformer update, language–visual fusion, SPG-based spatial position filtering, intermediate mask prediction, and frozen geometry feature injection. The predicted intermediate mask is further converted into q_s^ℓ and used as the historical localization signal for the next multimodal layer.

B.3 Cross-view Query Aggregation

CQA operates on the patch tokens output by the final multimodal layer. After removing the $R = 5$ register tokens from each view, the remaining features are concatenated as

$$\mathbf{F} \in \mathbb{R}^{NP\times D}. \tag{32}$$

We use $K_q = 20$ learnable shared queries and two query-decoder layers. Each layer contains 16-head cross-attention, query self-attention, and a feed-forward network.

For proposal-mask prediction, $\mathbf{F}$ is projected to mask features through

$$\mathbf{F}_m = H_m(\mathbf{F}), \qquad \mathbf{F}_m \in \mathbb{R}^{NP\times D}, \tag{33}$$

where $H_m(\cdot)$ denotes a two-layer MLP and $\mathbf{F}_m$ denotes the mask features used for proposal prediction. For the k-th shared query, the patch-level proposal logit at decoder layer ℓ is

$$S_{ks}^{\ell} = \frac{(\mathbf{u}_k^{\ell})^{\top}(\mathbf{F}_m)_s}{\sqrt{D}}, \tag{34}$$

where $\mathbf{u}_k^{\ell}$ denotes the k-th row of $\mathbf{Q}^{\ell}$ and S_{ks}^{ℓ} denotes the proposal logit of query k at patch position s.

The first query-decoder layer attends to all multi-view patches. For the second layer, the preceding proposal is converted into a binary attention constraint:

$$A_{ks}^{\ell} = \mathbb{I}\left[\sigma(S_{ks}^{\ell}) \geq 0.5\right], \tag{35}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function and A_{ks}^{ℓ} indicates whether patch position s is retained for query k. Positions below the threshold are masked in the subsequent cross-attention. If all positions of a query are suppressed, the attention mask is disabled for that query to avoid an empty attention set.

The query update follows the formulation in the main paper:

$$\mathbf{Q}^{\ell+1} = \mathrm{FFN}\big(\mathrm{SA}\big(\mathrm{CA}(\mathbf{Q}^{\ell}, \mathbf{F})\big)\big). \tag{36}$$

After the second query-decoder layer, the SCE-enhanced sentence representation $\mathbf{e}_{\mathrm{sent}}$ is used to weight the shared queries according to their relevance to the referring expression. In the implementation, $\mathbf{e}_{\mathrm{sent}}$ is represented by the first token of the SCE-enhanced language sequence. The query weights are computed as

$$\alpha_k = \mathrm{softmax}_k\left(\frac{(W_u\mathbf{u}_k)^{\top}W_s\mathbf{e}_{\mathrm{sent}}}{\sqrt{D}}\right), \tag{37}$$

consistent with the main-paper formulation.

The K_q query proposals are finally weighted by α_k to obtain the cross-view patch prediction, which is bilinearly upsampled to produce $\mathbf{M}_{\mathrm{CQA}}$. The aggregated prediction is then residually fused with the mask produced by the final multimodal layer:

$$\mathbf{M}_{\mathrm{final}} = \mathbf{M}^{12} + \eta\mathbf{M}_{\mathrm{CQA}}, \qquad \eta = \sigma(\beta), \tag{38}$$

where $\mathbf{M}^{12}$ denotes the mask predicted by the last multimodal layer, η denotes the learnable CQA fusion weight, and β denotes its pre-sigmoid scalar parameter. We initialize β to -2. This residual formulation enables CQA to exploit complementary target evidence across views while preserving the prediction produced by the main multimodal branch.

C More Experimental Results

We provide further analyses of training resolution, the CQA proposal threshold, auxiliary training objectives, and key hyperparameters. We first evaluate the complete MV-S^{2}D under the same 224×224 training resolution used by MVGGT to isolate the effect of training resolution. Importantly, regardless of the training resolution, all results in this paper follow the same test configuration as MVGGT: all eight sparse RGB views are used at a test resolution of 518×336, with the same dataset split and evaluation protocol.

Effect of training resolution. Our main model is trained at 448×448, whereas the original MVGGT uses a training resolution of 224×224. To rule out the possibility that the performance gain is mainly caused by higher-resolution training, we additionally train the complete MV-S^{2}D under the same 224×224 training resolution as MVGGT.

Even under this matched low-resolution setting, MV-S^{2}D improves Overall Global mIoU from 39.9 to 42.8, yielding a gain of 2.9 percentage points. More importantly, the improvement remains pronounced on the Multiple subset, which is most closely related to our disambiguation motivation.

Table 8: Performance comparison on MVRefer under different training resolutions. MV-S^{2}D (224) uses the same 224×224 training resolution as MVGGT, while MV-S^{2}D (448) denotes our main setting. All methods follow the same MVGGT evaluation protocol at a test resolution of 518×336. All values are mIoU (%).

	Hard ($\sim$ 40%)				Easy ($\sim$ 60%)				Unique ($\sim$ 19%)				Multiple ($\sim$ 81%)				Overall			
Method	Global	View	Pos	Neg	Global	View	Pos	Neg	Global	View	Pos	Neg	Global	View	Pos	Neg	Global	View	Pos	Neg
Two-stage	8.1	8.6	22.6	10.0	25.8	28.2	45.6	21.6	27.5	41.1	52.1	30.1	16.4	15.4	32.1	13.8	18.5	20.3	35.9	16.9
2D-Lift	6.4	15.0	25.3	11.9	25.4	24.1	45.1	12.3	31.5	30.9	47.2	19.2	14.5	17.9	34.8	10.4	17.8	20.4	37.2	12.1
MVGGT	24.4	67.3	31.6	78.0	50.1	70.6	52.2	81.2	65.2	82.6	64.5	90.4	33.8	66.1	39.0	77.4	39.9	69.3	44.0	79.9
MV-S^{2}D (224)	27.2	70.6	33.3	81.6	**53.2**	**73.3**	**54.1**	83.8	**67.3**	**84.1**	**65.8**	91.6	36.9	69.3	**41.0**	80.8	42.8	72.2	**45.8**	82.9
MV-S^{2}D (448)	**29.2**	**72.0**	**33.8**	**83.4**	52.5	73.1	52.1	**85.0**	67.1	**84.1**	64.7	**92.5**	**37.4**	**69.9**	40.0	**82.4**	**43.2**	**72.7**	44.8	**84.4**

Table 9: Performance comparison under the original ScanRefer protocol for MV-3DRES methods using sparse RGB inputs only. All methods follow the same MVGGT test configuration. Values are reported as percentages (%).

	Unique ($\sim$ 19%)			Multiple ($\sim$ 81%)			Overall		
Method	Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU
Two-stage	43.8	20.9	27.5	25.1	12.2	16.4	28.7	13.9	18.5
2D-Lift	51.6	26.0	31.5	21.5	5.6	14.5	27.3	9.6	17.8
MVGGT	83.6	74.5	65.2	49.2	33.5	33.8	55.9	41.5	39.9
MV-S^{2}D (224)	**86.9**	**77.5**	**67.3**	**52.4**	38.8	36.9	**59.1**	46.3	42.8
MV-S^{2}D (448)	85.4	77.0	67.1	51.2	**40.8**	**37.4**	57.9	**47.8**	**43.2**

Multiple mIoU increases from 33.8 to 36.9 (+3.1 points), while Acc@0.50 improves from 33.5 to 38.8 (+5.3 points). These gains show that, even without the additional visual details provided by higher-resolution training, the proposed semantic and spatial disambiguation mechanisms can more effectively distinguish among multiple candidate instances.

When the training resolution is further increased to 448×448, Multiple mIoU and Acc@0.50 further improve to 37.4 and 40.8, respectively. This suggests that higher-resolution training can provide additional benefits for mask quality, but is not the primary source of the disambiguation improvement. Overall, these results support our central motivation that the gains of MV-S^{2}D mainly arise from explicit semantic and spatial disambiguation, rather than from differences in training resolution.

Comparison under different IoU thresholds. We further compare Acc@0.25, Acc@0.50, and mIoU following the ScanRefer protocol. Table 9 includes only MV-3DRES methods that operate on sparse RGB inputs, avoiding direct comparison with traditional 3D-RES methods that rely on ground-truth point clouds. At the matched 224×224 training resolution, MV-S^{2}D improves Overall Acc@0.50 from 41.5 to **46.3** and Overall mIoU from 39.9 to **42.8** compared with MVGGT. On the challenging Multiple subset, Acc@0.50 also increases from 33.5 to 38.8. These results further demonstrate that the improvement remains substantial without a training-resolution advantage.

With the 448×448 main setting, Overall Acc@0.50 and mIoU are further improved to 47.8 and 43.2, respectively. Interestingly, the 224×224 model already achieves a higher Overall Acc@0.25 than the 448×448 model (59.1 vs. 57.9), indicating that higher-resolution training does not uniformly improve all metrics and mainly provides additional gains under stricter overlap requirements.

SCE supervision weight. We further analyze the effect of the category-supervision weight in SCE. As shown in Table 10, $\lambda_{\text{sce}} = 0.04$ achieves the best performance across Hard mIoU, Multiple mIoU, Acc@0.25, Acc@0.50, and Overall mIoU. Insufficient category supervision provides limited enhancement of target-category semantics, whereas overly strong supervision may suppress fine-grained attribute and relational cues in the referring expression. We therefore set $\lambda_{\text{sce}} = 0.04$.

Semantic category anchor visualization. To further analyze the category-aware representation learned by SCE, we visualize the matching distributions between referring expressions and semantic category anchors. As shown in Fig. 6, SCE assigns higher matching weights to category-consistent anchors. Expressions containing explicit category cues activate the corresponding anchors, while

Table 10: Ablation on the category-supervision weight λ_{sce} in SCE.

λ_{sce}	Hard mIoU	Multiple mIoU	Acc@0.25	Acc@0.50	mIoU
0	26.9	35.9	57.7	44.6	41.8
0.02	26.0	35.9	57.9	45.1	41.9
0.04	**27.2**	**36.9**	**59.1**	**46.3**	**42.8**
0.08	26.7	36.0	57.8	45.1	41.9

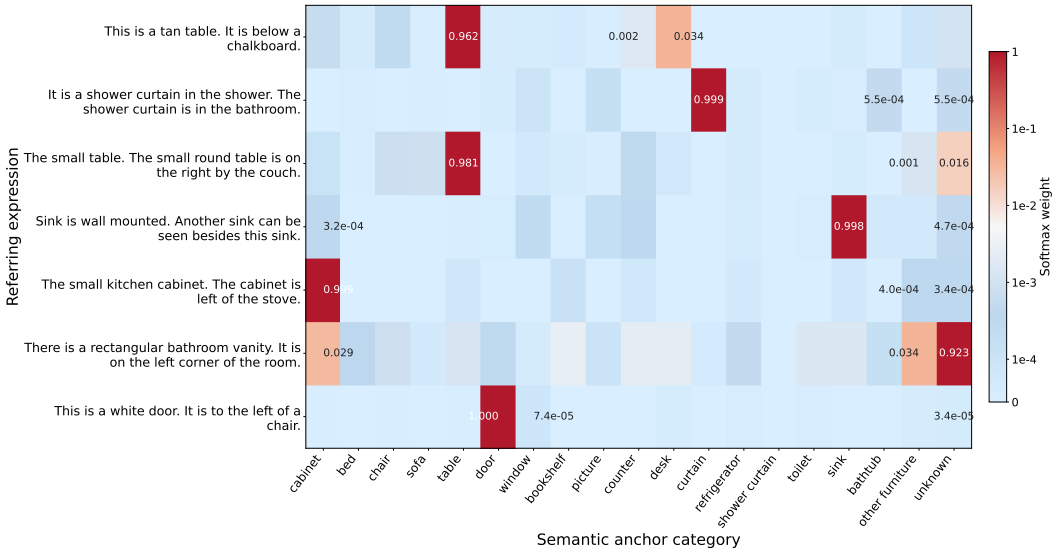


Figure 6: Visualization of semantic category anchor matching in SCE. The heatmap illustrates the matching weights between referring expressions and learnable category anchors, where each row corresponds to a referring expression and each column represents a category anchor.

Table 11: Ablation on auxiliary training objectives.

Setting	Hard mIoU	Multiple mIoU	Acc@0.25	Acc@0.50	mIoU
w/o Background Suppression	26.8	36.2	58.2	45.2	42.1
w/o Patch Contrastive Loss	26.7	35.9	57.8	45.1	41.8
Full Objective	**27.2**	**36.9**	**59.1**	**46.3**	**42.8**

ambiguous descriptions maintain softer distributions over related categories. This indicates that SCE can capture target-category semantics before multimodal interaction.

Auxiliary training objectives. We further investigate the two auxiliary objectives used for CQA. As shown in Table 11, removing the background-suppression loss decreases Overall mIoU from 42.8 to 42.1, while removing the patch-level contrastive loss further reduces it to 41.8. The latter also causes larger degradation on the Hard and Multiple subsets. The background-suppression loss reduces erroneous foreground responses in non-target regions, whereas the patch-level contrastive objective improves feature separability between target and background. Their combination provides more stable optimization and better final segmentation performance.

CQA candidate-mask threshold and shared queries. Finally, we analyze two key hyperparameters in CQA: the candidate-mask threshold θ_q in the second query-decoder layer and the number of shared queries K_q. As shown in Table 12, $\theta_q = 0.5$ achieves the highest Overall mIoU of 42.8 and Acc@0.25 of 59.1, together with a tied-best Acc@0.50 of 46.3. A lower threshold retains more low-confidence background positions, whereas an overly high threshold may suppress useful target evidence.

Table 12: Ablation on the candidate-mask threshold θ_q in the second CQA layer.

θ_q	Hard mIoU	Multiple mIoU	Acc@0.25	Acc@0.50	mIoU
0.4	27.1	36.4	58.5	45.2	42.2
0.5	27.2	**36.9**	**59.1**	**46.3**	**42.8**
0.6	**27.3**	36.6	58.1	**46.3**	42.4
0.7	27.2	36.5	58.3	45.4	42.3

Table 13: Ablation on the number of shared queries K_q in CQA.

K_q	Hard mIoU	Multiple mIoU	Acc@0.25	Acc@0.50	mIoU
10	26.7	36.0	58.2	45.0	41.9
20	**27.2**	**36.9**	**59.1**	**46.3**	**42.8**
30	27.2	36.8	58.6	45.3	42.5
40	26.9	36.1	57.8	45.6	42.1

For the number of shared queries, increasing K_q from 10 to 20 improves Overall mIoU from 41.9 to 42.8. Further increasing the number of queries to 30 or 40 yields diminishing returns and eventually degrades performance. We therefore set $\theta_q = 0.5$ and $K_q = 20$.

D More Visual Comparison Results

We further provide additional qualitative comparisons on MVRefer. These examples mainly contain multiple same-category or visually similar candidate instances, where the referring expressions rely on category information, attributes, and spatial relations to identify the target. MVGGT may be distracted by ambiguous candidates, whereas MV-S^{2}D leverages semantic and spatial cues to reduce referential ambiguity and produce more accurate target segmentation.

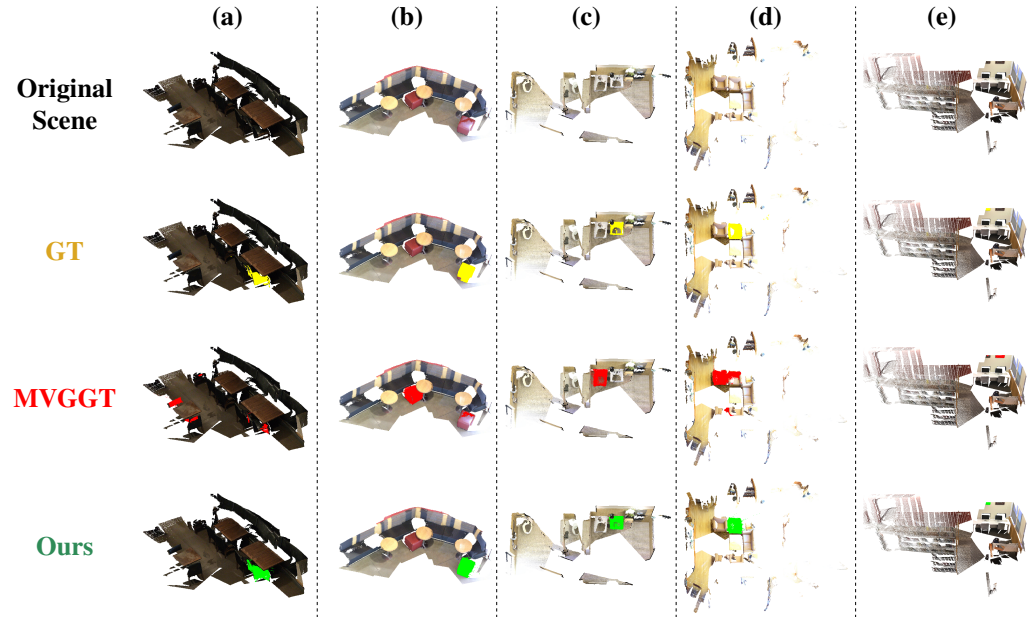


Figure 7: (1) Additional qualitative comparisons on MVRefer under semantic and spatial ambiguity.

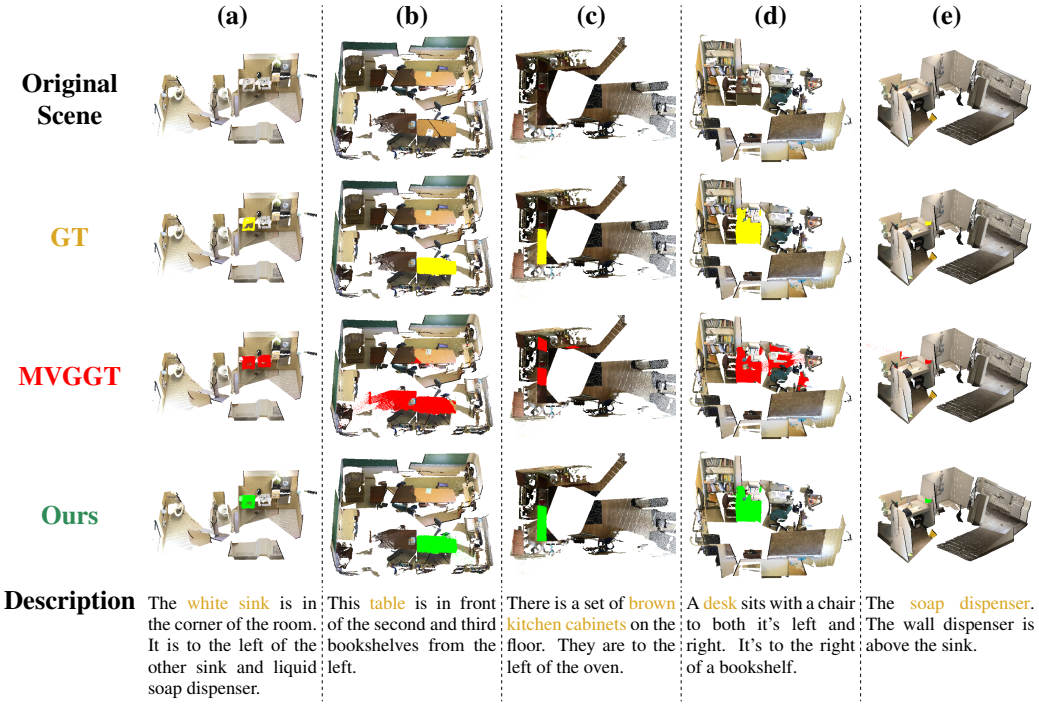


Figure 8: (2) Additional qualitative comparisons on MVRefer under semantic and spatial ambiguity.

E Inference Efficiency

To analyze the computational overhead introduced by MV-S^{2}D, we compare the inference efficiency of MVGGT and MV-S^{2}D under the same hardware and evaluation settings. Both methods are evaluated on an NVIDIA RTX PRO 6000 GPU with the same input resolution and testing protocol.

Table 14: Inference efficiency comparison under the same hardware and evaluation settings. All methods are evaluated with a test resolution of 518×336.

Method	Test Resolution	Overall mIoU	Inference Time (s/batch)	GPU Memory (GB)
MVGGT	518×336	39.9	3.3687	14.1
MV-S^{2}D	518×336	43.2	3.8885	16.2