This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MVGGT: Multimodal Visual Geometry Grounded Transformer for Multiview 3D Referring Expression Segmentation

Changli Wu[1,2,†], Haodong Wang[1,†], Jiayi Ji[1,*], Yutian Yao[5],
Chunsai Du[4], Jihua Kang[4], Yanwei Fu[2,3], Liujuan Cao[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
[2]Shanghai Innovation Institute [3]Fudan University [4]ByteDance
[5]Tianjin University of Science and Technology

{wuchangli, wahadon}@stu.xmu.edu.cn, jjyxmu@gmail.com, 22201316@mail.tust.edu.cn,
{duchunsai, kangjihua}@bytedance.com, yanweifu@fudan.edu.cn, caoliujuan@xmu.edu.cn

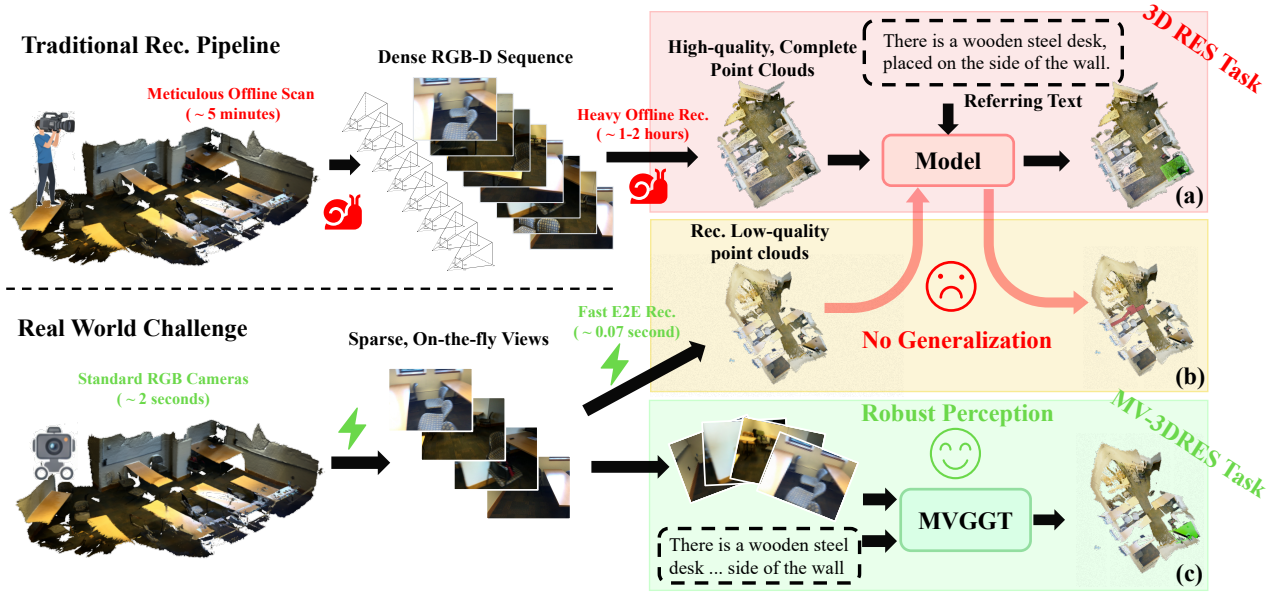


Figure 1. **The Reality Gap: From Idealized 3D RES to Real-World MV-3DRES. (a)** Traditional 3D RES depends on dense, high-quality point clouds produced by slow offline scanning and heavy reconstruction. **(b)** Applied to sparse, low-quality point clouds from real-world RGB views, these models fail to generalize. **(c)** We introduce MV-3DRES, which uses sparse multi-view RGB inputs and text to achieve robust joint reconstruction and perception, enabled by our MVGGT model.

Abstract

Most existing 3D referring expression segmentation (3DRES) methods rely on dense, high-quality point clouds, while real-world agents such as robots and mobile phones operate with only a few sparse RGB views and strict latency constraints. We introduce Multi-view 3D Referring Expression Segmentation (MV-3DRES), where the model must recover scene structure and segment the referred object directly from sparse multi-view images. Traditional two-stage pipelines, which first reconstruct a point cloud and then perform segmentation, often yield low-quality geometry, pro-

[†]Equal Contribution.
[*]Corresponding Author.

duce coarse or degraded target regions, and run slowly. We propose the Multimodal Visual Geometry Grounded Transformer (MVGGT), an efficient end-to-end framework that integrates language information into sparse-view geometric reasoning through a dual-branch design. Training in this setting exposes a critical optimization barrier, termed Foreground Gradient Dilution (FGD), where sparse 3D signals lead to weak supervision. To resolve this, we introduce Per-view No-target Suppression Optimization (PVSO), which provides stronger and more balanced gradients across views, enabling stable and efficient learning. To support consistent evaluation, we build MVRefer, a benchmark that defines standardized settings and metrics for MV-3DRES. Experiments show that MVGGT establishes the first strong baseline and achieves both high accuracy and fast inference, outperforming existing alternatives. The code is available at https://mvggt.github.io/.

1. Introduction

Grounding natural language in 3D physical scenes is fundamental to embodied AI. A prominent formulation is 3D referring expression segmentation (3DRES), where a model segments an object in a 3D scene given a textual description. Although recent methods [1, 3, 24, 59–61, 66, 68] have achieved strong results, they are built upon a rarely questioned assumption: the availability of dense, complete, and reliable point clouds. Such point clouds typically require LIDAR sensors or lengthy RGB-D SLAM pipelines like BundleFusion [7], which demand deliberate scanning and heavy offline processing. This assumption stands in stark contrast to real-world agents—robots, AR glasses, mobile devices—that perceive environments through only a few casually captured RGB views.

In real settings, high-fidelity geometry is the exception. Sparse multi-view images produce 3D reconstructions that are noisy, incomplete, and often ambiguous (Figure 1(b)). Existing 3DRES models, trained on idealized point clouds, collapse under such inputs. This exposes a fundamental limitation: current 3DRES is sensor-privileged and misaligned with the actual sensing conditions of embodied systems. It motivates a central question: **How can we achieve language-grounded 3D perception when complete geometry is no longer given but must be inferred from sparse, inconsistent views?**

We address this by introducing Multi-view 3D Referring Expression Segmentation (MV-3DRES), a new setting where the model must jointly reconstruct the scene and segment the referred object directly from sparse RGB views (Figure 1(c)). MV-3DRES is inherently challenging: the model must reason over missing structure, integrate information across misaligned viewpoints, and resolve linguistic ambiguities without access to dense 3D input.

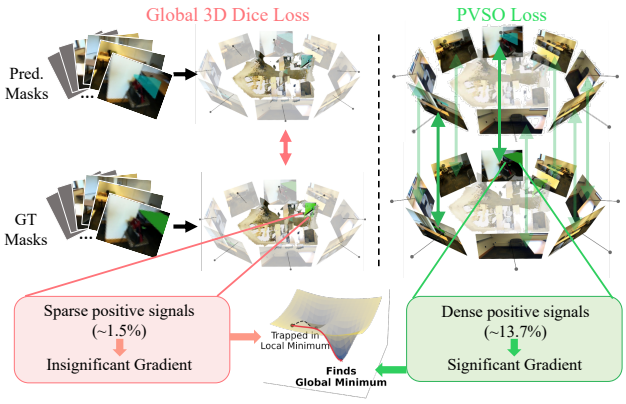


Figure 2. Illustration of Foreground Gradient Dilution Problem of Global 3D DICE loss and Per-view No-Target Suppression Optimization.

Conventional pipelines fail in this regime. Purely 2D methods cannot enforce global 3D consistency, since they operate on isolated views and cannot resolve depth ordering, occlusion relationships, or spatial relations such as "in front of" or "on the left." As a result, back-projecting their per-view masks produces fragmented or conflicting 3D predictions. Two-stage "reconstruct-then-segment" pipelines face a different failure mode: sparse inputs yield point clouds that are noisy, incomplete, and structurally distorted, making it difficult for 3DRES models to recover full object extents. Moreover, running a full reconstruction before segmentation incurs substantial latency, limiting practical deployment.

To this end, we propose the Multimodal Visual Geometry Grounded Transformer (MVGGT), the first end-to-end architecture designed specifically for MV-3DRES. MVGGT adopts a dual-branch paradigm: a frozen geometric branch provides camera poses, depth cues, and a coarse structural scaffold, while a multimodal branch injects linguistic cues into sparse-view visual features through cross-view, cross-modal attention. This design embodies a key conceptual shift: language is intertwined with geometric reasoning from the start, enabling it to guide evidence aggregation and scene disambiguation long before a complete 3D representation exists.

However, sparse-view learning introduces a fundamental optimization challenge. Sparse multi-view reconstruction produces point clouds in which the target instance is represented by only a very small number of scattered points, far fewer than in the dense point clouds used by conventional 3DRES methods. Under such extreme foreground sparsity, standard 3D losses such as Dice become ineffective: gradients from the target region are overwhelmed by background points, leading to Foreground Gradient Dilution (FGD) and causing the optimization to stagnate in early

training—gradients are too small to escape poor local minima, as illustrated in Figure 2. The problem is exacerbated by view-dependent visibility—some views contain clear target evidence, while others provide almost none—making uniform 3D supervision unstable and noisy. To mitigate FGD, we introduce Per-view No-target Suppression Optimization (PVSO), which shifts supervision back to 2D view space where the target occupies a larger and more reliable area. This per-view formulation amplifies meaningful gradients from informative views and suppresses misleading signals from target-absent views, resulting in significantly more stable and effective training.

Finally, to standardize evaluation, we construct MVRefer, the first benchmark defining settings, metrics, and data protocol for MV-3DRES. Extensive experiments show that MVGGT provides a strong baseline and significantly outperforms existing alternatives. Taken together, our contributions are fourfold:

- We identify and formalize MV-3DRES, a new problem setting that aligns 3D grounding with realistic sensing conditions.
- We propose MVGGT, a novel dual-branch architecture unifying geometric scaffolding with cross-view, language-aware perception.
- We analyze and address the Foreground Gradient Dilution challenge via PVSO, offering a principled optimization strategy tailored for sparse 3D supervision.
- We construct the MVRefer benchmark, defining standardized settings and metrics for MV-3DRES and providing the first strong baseline.

2. Related Work

2.1. Traditional 3D Referring Segmentation

3D grounding [12, 14, 17, 18, 25, 27, 29, 30, 38, 43, 46, 49, 67] aims to locate a specific object in a 3D scene based on a unique natural language description [1, 3],which is part of vision-language tasks [13, 15, 16, 33, 37, 62]. Following this, the 3D-RES (3D Referring Segmentation) task [4, 5, 8, 21–23, 26, 34, 41] aims to segment a specific object within a point cloud based on a textual query. The field has evolved from foundational two-stage paradigms [24, 47, 64] (relying on object proposals and language matching) to recent end-to-end architectures [19, 20, 36, 39, 43, 47, 59–61] demonstrating high efficacy through advanced cross-modal fusion. Despite this progress, the sub-field shares a critical limitation: the requirement for high-quality, dense point clouds. This expensive geometric data is inaccessible to many real-world embodied agents that must rely on sparse, online RGB captures.

2.2. Multi-View Feed-forward Reconstruction

Reconstructing 3D geometry from multi-view RGB images provides a practical solution to the input ambiguity in sparse-view settings. Early feed-forward approaches such as DUSt3R and MASt3R [31, 55] introduced coupled scene representations but required heavy post-processing or integration with classical SfM/SLAM pipelines [10, 11, 44, 45] for unconstrained reconstruction. Later works improved efficiency and stability by replacing classical optimization with Transformer-based latent state propagation, as demonstrated by Spann3R, CUT3R, and MUSt3R [2, 52, 54]. Streaming models like WinT3R [58] further enabled real-time performance via sliding-window processing and global camera token pooling.

More recent architectures—VGGT and its successors [9, 48, 51, 53, 56]—adopt alternating-attention designs to achieve robust generalized reconstruction, while extensions address semantic and multi-task perception [32, 50, 57, 63]. This progress culminates in universal backbones such as MapAnything [28], which produce fully factored, metric-aware scene representations.

3. MV-3DRES Task and MVRefer Benchmark

3.1. Task Formulation

We formalize the Multi-view 3D Referring Segmentation (MV-3DRES) task to align 3D language grounding with the sensing constraints of real-world agents. Instead of assuming access to pre-constructed dense point clouds, the model operates directly on sparse multi-view RGB images.

Given a set of N RGB views $I = \{I_i\}_{i=1}^{N}$ and a natural-language referring expression T, the goal is to learn a function

$$f : (I, T) \rightarrow (S', M), \tag{1}$$

where $S' \in \mathbb{R}^{K\times 3}$ denotes the reconstructed 3D point cloud containing K points, and $M \in \{0, 1\}^K$ is the corresponding 3D binary mask marking the points that belong to the object referred to by T. The model must infer both geometry and semantics from the same sparse observations, without any ground-truth 3D input at inference time.

This formulation introduces challenges not present in standard 3DRES. Sparse multi-view observations generate incomplete and noisy geometry, forcing the model to couple reconstruction and grounding. Spatial relations described in language, such as “on the left of the chair,” must be resolved across viewpoints with inconsistent visibility. Moreover, the target object often occupies only a small portion of the available views, yielding severe foreground sparsity and weak supervisory signals, which we later characterize as Foreground Gradient Dilution (FGD).

3.2. The MVRefer Benchmark

To support systematic evaluation of MV-3DRES, we construct MVRefer, a benchmark built upon ScanRefer [3] and the underlying ScanNet sequences [6]. MVRefer is designed to emulate how an embodied agent perceives a scene through a limited number of casual views.

3.2.1. Benchmark Setting

For each language–object pair in ScanRefer [3], we sample $N = 8$ RGB frames from the raw ScanNet video stream [6] at uniform temporal intervals to approximate sparse, on-the-fly observations. Sparse sampling creates a solvability issue: the target may be absent from all selected frames. To ensure each sample remains resolvable, we perform a visibility validation step. If none of the initial eight images contain the target, we replace one no-target frame with a randomly chosen target-visible frame. This guarantees at least one positive view while naturally preserving a high proportion of no-target views, maintaining the difficulty inherent to sparse-view grounding.

3.2.2. Evaluation Metrics and Splits

Evaluating MV-3DRES requires metrics that disentangle grounding quality from reconstruction quality. Since both outputs in (S', M) are jointly predicted from sparse inputs, reconstruction errors can obscure the model's true grounding ability.

Traditional 3D Metric. We report global 3D mean IoU,

$$\text{mIoU}_{\text{global}} = \text{IoU}(M, M^*), \tag{2}$$

where M^* denotes the ground-truth mask projected onto the reconstructed point cloud S'. Although standard, mIoUglobal entangles segmentation performance with the fidelity of S', making it insufficient as the primary diagnostic measure.

Multi-view Diagnostic Metrics. To isolate grounding behaviors, we reproject the predicted 3D mask M into each view using the known camera intrinsics and extrinsics. Let $P_i(M)$ denote the projected 2D mask for view i, and $P_i(M^*)$ its ground-truth counterpart. We compute:

$$\text{mIoU}_{\text{view}} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}\left(P_i(M), P_i(M^*)\right), \tag{3}$$

$$\text{mIoU}_{\text{pos}} = \frac{1}{|\mathcal{V}^+|} \sum_{i \in \mathcal{V}^+} \text{IoU}\left(P_i(M), P_i(M^*)\right), \tag{4}$$

$$\text{mIoU}_{\text{neg}} = \frac{1}{|\mathcal{V}^-|} \sum_{i \in \mathcal{V}^-} \text{IoU}\left(P_i(M), P_i(M^*)\right), \tag{5}$$

where $\mathcal{V}^+$ and $\mathcal{V}^-$ denote the sets of target-visible and no-target views, respectively. These metrics shed light on grounding precision (mIoU_{pos}) and suppression ability (mIoU_{neg}), both of which are essential for robust performance under sparse supervision.

Difficulty Splits. To evaluate robustness under varying signal sparsity, we define two difficulty splits based on the target's 2D pixel ratio. A sample is categorized as *hard* if the target occupies less than 5% of pixels in all its visible views, and *easy* if at least one view contains at least 5% target pixels. This separation allows us to isolate performance differences arising from the strength of view-specific supervision.

4. Method

We introduce the **Multimodal Visual Geometry Grounded Transformer (MVGGT)**, an end-to-end dual-branch framework tailored for the MV-3DRES task, as shown in Figure 3.

4.1. The proposed MVGGT

MVGGT is designed to jointly recover 3D geometry and perform language-conditioned segmentation from sparse views. The separation into two branches allows the model to exploit a stable geometric scaffold while learning multimodal representations aligned with the text query.

Inputs and Encoders. Given N input images $I = \{I_i\}_{i=1}^{N}$ and a referring expression T, each image is encoded by a frame-wise ViT, yielding patch embeddings $F_i^{\text{vis}} \in \mathbb{R}^{P \times D}$, where P is the number of patches and D the feature dimension. The text is tokenized and processed by a language encoder to produce word embeddings $F^{\text{lang}} \in \mathbb{R}^{W \times D}$, where W denotes token count.

Frozen Reconstruction Branch. The reconstruction branch is a geometry-aware transformer with L blocks. Each block alternates between frame-level self-attention and global cross-view attention, progressively building view-consistent structural cues. Let F_ℓ^{geo} denote the features at block $\ell \in \{1, \ldots, L\}$. These features are fed to a reconstruction decoder that predicts camera poses and depth maps, which are back-projected into a coarse point cloud S'. All parameters in this branch remain frozen, ensuring a stable geometric prior across training and removing the need to re-learn 3D geometry from sparse images.

Trainable Multimodal Branch. The multimodal branch contains $L_{\text{multi}} = L/3$ transformer blocks. Its goal is to fuse geometric cues with text-conditioned visual features. Since

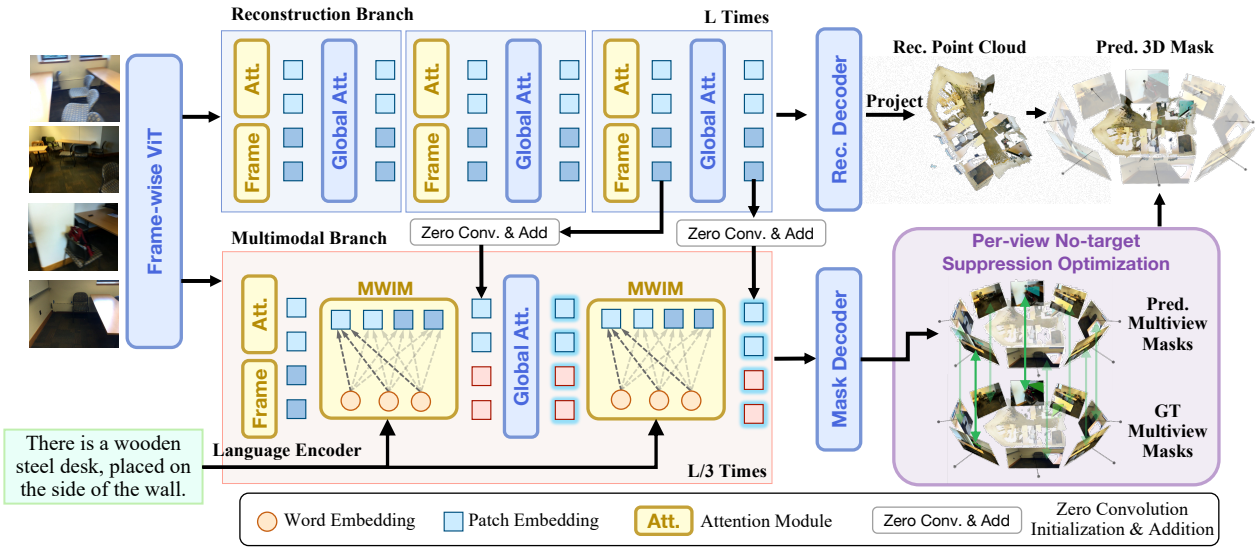


Figure 3. Architecture of MVGGT, which comprises a frozen Reconstruction Branch that establishes geometric structure and a trainable Multimodal Branch that integrates language into sparse-view visual reasoning.

the two branches have different depths, we align their interactions by injecting geometric features from the final $L/3$ blocks of the reconstruction branch into all L_{multi} blocks of the multimodal branch.

Geometric Injection. The multimodal branch contains $L_{\text{multi}} = L/3$ blocks, and each of them receives geometric guidance from the *final* L_{multi} blocks of the reconstruction branch. Concretely, for the l'-th multimodal block ($l' = 1, \ldots, L_{\text{multi}}$), we take the geometric feature F_l^{geo} from the l-th reconstruction block, where l runs over the last L_{multi} layers in order (i.e., the reconstruction branch's $(L-L_{\text{multi}}+1)$-th layer provides geometry to the first multimodal block, the next layer to the second, and so on).

These geometric features are passed through a zero-initialized 1×1 convolution $\mathcal{Z}$ [65], which projects them into the multimodal feature space. The input to the l'-th multimodal block is then

$$F_{l'}^{\text{in}} = F_{l'-1}^{\text{out}} + \mathcal{Z}(F_l^{\text{geo}}), \tag{6}$$

where $F_{l'-1}^{\text{out}}$ is the output of the preceding visual attention and cross attention. It allows the multimodal branch to progressively incorporate higher-level geometric cues without perturbing the pretrained reconstruction backbone. $F_{l'}^{\text{in}}$ is then fed into the visual attention module to get $F_{l'}^{\text{vis}}$.

Language Injection. Within each multimodal block, language injection is implemented through a standard cross-attention layer. Let $F_{l'}^{\text{in}}$ be the visual tokens entering block l'. Query, key, and value projections are defined as

$$Q = F_{l'}^{\text{vis}} W_Q, \quad K = F^{\text{lang}} W_K, \quad V = F^{\text{lang}} W_V, \tag{7}$$

with W_Q, W_K, and $W_V \in \mathbb{R}^{D \times D}$ learnable matrices. Attention is computed as

$$F_{l'}^{\text{out}} = \text{softmax}\left(\frac{QK^\top}{\sqrt{D}}\right) V, \tag{8}$$

and added back through a residual path. This multi-level injection allows text cues to guide feature aggregation across views.

Decoders and Outputs. Finally, the multimodal features $F_{L_{\text{multi}}}^{\text{multi}}$ are decoded into per-view masks $\{M_i\}_{i=1}^N$. Using the depths and camera parameters from the frozen reconstruction branch, these 2D masks are back-projected and aggregated on the reconstructed point cloud S' to obtain the final 3D mask M. The resulting multi-view predictions serve as the supervision targets in PVSO (Section 4.3).

4.2. Foreground Gradient Dilution

Training under the MV-3DRES setting is fundamentally hindered by the extreme sparsity of foreground points in the reconstructed 3D space. Let the Dice loss for 3D segmentation be

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2I}{U}, \quad I = \sum_j p_j g_j, \quad U = \sum_j p_j + \sum_j g_j, \tag{9}$$

where p_j is the predicted probability at point j and $g_j \in \{0, 1\}$ denotes ground-truth labels. The gradient with respect to p_j is

$$\frac{\partial \mathcal{L}_{\text{Dice}}}{\partial p_j} = \frac{2(I - g_j U)}{U^2}. \tag{10}$$

During early training, predictions remain small and diffuse, yielding $I \approx 0$. In MV-3DRES, reconstructed point clouds from sparse views are large (often 10^6–10^7 points) while the target instance typically occupies less than 2% of them. Consequently, the union term U is dominated by background, inflating by several orders of magnitude. For a foreground point ($g_j = 1$),

$$\left.\frac{\partial \mathcal{L}_{\text{Dice}}}{\partial p_j}\right|_{g_j=1} \approx -\frac{2}{U}, \tag{11}$$

whose magnitude becomes extremely small when U is large. Empirically, gradients fall to 10^{-9}–10^{-11}, far below the scale needed to drive meaningful updates. Although foreground points are present, their contribution to optimization becomes negligible, leading to stalled convergence. We refer to this failure mode as *Foreground Gradient Dilution (FGD)*.

4.3. Per-view No-Target Suppression Optimization

To mitigate FGD, we propose **Per-view No-Target Suppression Optimization (PVSO)**, a view-wise supervision strategy that shifts early learning signals from sparse 3D space to the denser 2D image domain. This modification significantly reduces the imbalance between foreground and background, thereby amplifying effective gradients.

Positive-aware Sampling. Given the view set $\mathcal{V}$ of a scene, let $\mathcal{V}_t$ and $\mathcal{V}_n$ denote target-visible and no-target views, respectively. PVSO samples a subset $\mathcal{V}'$ while enforcing a minimum foreground-view ratio

$$\rho_t = \frac{|\mathcal{V}_t|}{|\mathcal{V}'|}, \tag{12}$$

ensuring that each batch contains sufficient positive evidence. This prevents the optimization from collapsing to trivial background predictions.

2D Gradient Concentration. For each sampled view, the predicted 3D mask is projected onto the image plane, and a 2D Dice loss is applied. Since foreground regions typically occupy 10–15% of pixels in visible views—far larger than their $< 2\%$ proportion in 3D point clouds—the 2D Dice denominator U_{2D} becomes substantially smaller:

$$U_{\text{2D}} \ll U_{\text{3D}}. \tag{13}$$

Given that the Dice gradient scales as $\mathcal{O}(1/U)$, the per-view supervision yields foreground gradients that are 1–3 orders of magnitude larger than in 3D. This concentrated 2D supervision strengthens early training signals.

Suppression of No-Target Views. No-target views often far outnumber target-visible ones. To avoid overwhelming the loss with trivial negatives, PVSO normalizes their contribution using

$$w_s = \frac{1}{|\mathcal{V}_n|}. \tag{14}$$

The complete PVSO objective is

$$L_{\text{PVSO}} = \frac{1}{|\mathcal{V}_t| + 1}\Big(\sum_{i \in \mathcal{V}_t} L_{\text{Dice}}(m_i, M_i^{\text{gt}}) + w_s \sum_{j \in \mathcal{V}_n} L_{\text{Dice}}(m_j, \mathbf{0}) \Big), \tag{15}$$

where m_i is the predicted 2D mask for view i and M_i^{gt} is the corresponding ground-truth mask (empty for no-target views).

Joint Objective. PVSO complements 3D supervision by providing dense, stable gradients during early training. The complete objective is

$$L_{\text{total}} = L_{\text{BCE}} + \lambda_p L_{\text{PVSO}}, \tag{16}$$

where λ_p balances 2D and 3D signals. This formulation alleviates foreground gradient dilution and yields robust multimodal 3D grounding from sparse views.

5. Experiments

5.1. Implementation Details and Setup

MVGGT Configuration. We adopt the Pi3 [56] reconstruction backbone (36 blocks), which remains frozen throughout training. We use a frozen Roberta model [40] as the language encoder. The multimodal branch contains $L_{\text{multi}} = 12$ blocks and is optimized end-to-end. Training uses AdamW [42] with a 1×10^{-4} learning rate, batch size of 16, and 30 epochs on a single NVIDIA 4090 GPU. The PVSO weight λ_p is fixed to 1.

Dataset and Metrics. All evaluations are conducted on the proposed **MVRefer** benchmark. We follow standard ScanRefer train/validation splits [3]. Alongside $\text{mIoU}_{\text{global}}$, we report all diagnostic view-level metrics from Section 3.2.2, with special focus on the **Hard** and **Easy** subsets, which most directly reflect the foreground sparsity challenge underlying FGD.

5.2. Main Results

We evaluate MVGGT against two MV-3DRES baselines: (1) **2D-Lift**, which projects ReferDINO [35] masks into 3D via Pi3 [56]; and (2) **two-stage**, a sequential pipeline of Pi3 [56] reconstruction and LESS [39] segmentation. As shown in Table 1, MVGGT consistently leads across all

Table 1. Performance comparison on the MVRefer benchmark.Metrics are all mIoU under different categories.

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	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
Ours	**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**

Table 2. The results of traditional 3D-RES and MV-3DRES tasks under original ScanRefer setting.

Method	Unique (∼19%)			Multiple (∼81%)			Overall		
	Acc@25	Acc@50	mIoU	Acc@25	Acc@50	mIoU	Acc@25	Acc@50	mIoU
Traditional 3D-RES (with ground-truth point clouds)									
TGNN [24]	69.3	57.8	50.7	31.2	26.6	23.6	38.6	32.7	28.8
3D-STMN [61]	89.3	84.0	74.5	46.2	29.2	31.1	54.6	39.8	39.5
SegPoint [20]	-	-	-	-	-	-	-	-	41.7
Reason3D [23]	88.4	84.2	74.6	50.5	31.7	34.1	57.9	41.9	42.0
RG-SAN [59]	89.2	84.3	74.5	55.0	35.4	37.4	61.7	44.9	44.6
LESS [39]	-	-	-	-	-	-	53.2	29.9	33.7
3D-LLaVA [8]	-	-	-	-	-	-	-	-	43.3
MV-3DRES (sparse RGB inputs only)									
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
Ours	**83.6**	**74.5**	**65.2**	**49.2**	**33.5**	**33.8**	**55.9**	**41.5**	**39.9**

difficulty levels. On **Hard**, it achieves 24.4 global mIoU, surpassing two-stage and 2D-Lift by 16.3 and 18.0, respectively, with a 52.3 gain in view mIoU. This pattern holds for **Easy** (50.1 global; 70.6 view mIoU). Overall, MVGGT attains 39.9 global mIoU, exceeding the strongest baseline by 22.1. These gains validate MVGGT's robustness to sparsity and the effectiveness of PVSO in mitigating FGD.

Table 2 evaluates MVGGT with traditional 3D-RES and recent MV-3DRES baselines under the standard ScanRefer protocol. Despite using only sparse RGB, MVGGT achieves 65.2 mIoU on *Unique*, significantly closing the gap with full 3D-resourced methods. In *Multiple*, MVGGT reaches 33.8 mIoU, outperforming two-stage and 2D-Lift by 17.4 and 19.3. These results demonstrate that MVGGT enables reliable 3D grounding without ground-truth point clouds.

5.3. Ablation Studies

We conduct comprehensive ablation studies to validate the effectiveness of each proposed component. All experiments are performed on the MVRefer benchmark.

Impact of Core Components. Table 3 evaluates core components on the MVRefer benchmark. Removing both MVGGT and PVSO causes a significant performance drop, confirming the difficulty of sparse-view 3D grounding. Individually, PVSO improves $\mathrm{mIoU}_{\mathrm{global}}$ ($26.9 \rightarrow 32.0$) and $\mathrm{mIoU}_{\mathrm{view}}$ ($41.1 \rightarrow 47.5$) by mitigating Foreground Gradient Dilution. MVGGT enhances robustness, particularly on hard scenes ($19.0 \rightarrow 24.4$ $\mathrm{mIoU}_{\mathrm{global}}$), by guiding geometric reasoning with linguistic cues. The full model achieves 39.9 $\mathrm{mIoU}_{\mathrm{global}}$ and 69.3 $\mathrm{mIoU}_{\mathrm{view}}$, validating the synergy between our multimodal design and optimization strategy.

Table 3. Ablation studies on the core components.

PVSO	MVGGT	Easy ↑		Hard ↑		Overall ↑	
		$\mathrm{mIoU}_{\mathrm{global}}$	$\mathrm{mIoU}_{\mathrm{view}}$	$\mathrm{mIoU}_{\mathrm{global}}$	$\mathrm{mIoU}_{\mathrm{view}}$	$\mathrm{mIoU}_{\mathrm{global}}$	$\mathrm{mIoU}_{\mathrm{view}}$
2D-Lift		25.4	24.1	6.4	15.0	17.8	20.4
×	×	36.3	43.0	12.9	38.5	26.9	41.1
✓	×	40.7	48.9	19.0	45.4	32.0	47.5
✓	✓	**50.1**	**70.6**	**24.4**	**67.3**	**39.9**	**69.3**

PVSO Component Analysis. Table 4 evaluates PVSO variants. Without no-target suppression, random sampling yields unstable performance. Enabling suppression improves $\mathrm{mIoU}_{\mathrm{global}}$ ($32.4 \rightarrow 36.7$) by filtering misleading gradients from target-absent views, while hybrid sampling ensures sufficient per-batch positive evidence. A 0.5 no-target ratio achieves the best balance (39.9 $\mathrm{mIoU}_{\mathrm{global}}$, 69.3 $\mathrm{mIoU}_{\mathrm{view}}$); extreme ratios either underexploit negative pairs or drown out positives. These results confirm that stable sparse-view learning requires weighting view contributions by their actual target visibility.

Table 4. Ablation on PVSO components.

Sampling Strategy	No-Target Ratio	No-Target Suppression	Easy ↑		Hard ↑		Overall ↑	
			$\mathrm{mIoU}_{\mathrm{global}}$	$\mathrm{mIoU}_{\mathrm{view}}$	$\mathrm{mIoU}_{\mathrm{global}}$	$\mathrm{mIoU}_{\mathrm{view}}$	$\mathrm{mIoU}_{\mathrm{global}}$	$\mathrm{mIoU}_{\mathrm{view}}$
Random	-	×	40.8	65.1	19.7	66.5	32.4	65.6
Random	-	✓	46.2	55.8	22.4	49.3	36.7	53.2
Hybrid	0	✓	42.4	24.3	11.2	10.3	29.9	18.7
Hybrid	0.25	✓	46.5	65.2	19.4	59.6	35.7	63.0
Hybrid	0.5	✓	**50.1**	**70.6**	**24.4**	**67.3**	**39.9**	**69.3**
Hybrid	0.75	✓	30.3	64.0	18.7	72.0	25.7	67.2

MVGGT Fusion Architecture. Table 5 evaluates multimodal fusion placement. Early fusion performs weakest, as injecting language before geometric features form can disrupt structural reasoning. Middle fusion shows modest gains, while late fusion yields the best results (39.9 $\mathrm{mIoU}_{\mathrm{global}}$; 69.3 $\mathrm{mIoU}_{\mathrm{view}}$). This trend suggests that spatial perception should be established first, with language providing later refinement. Such an order ensures more stable and discriminative cross-view representations.

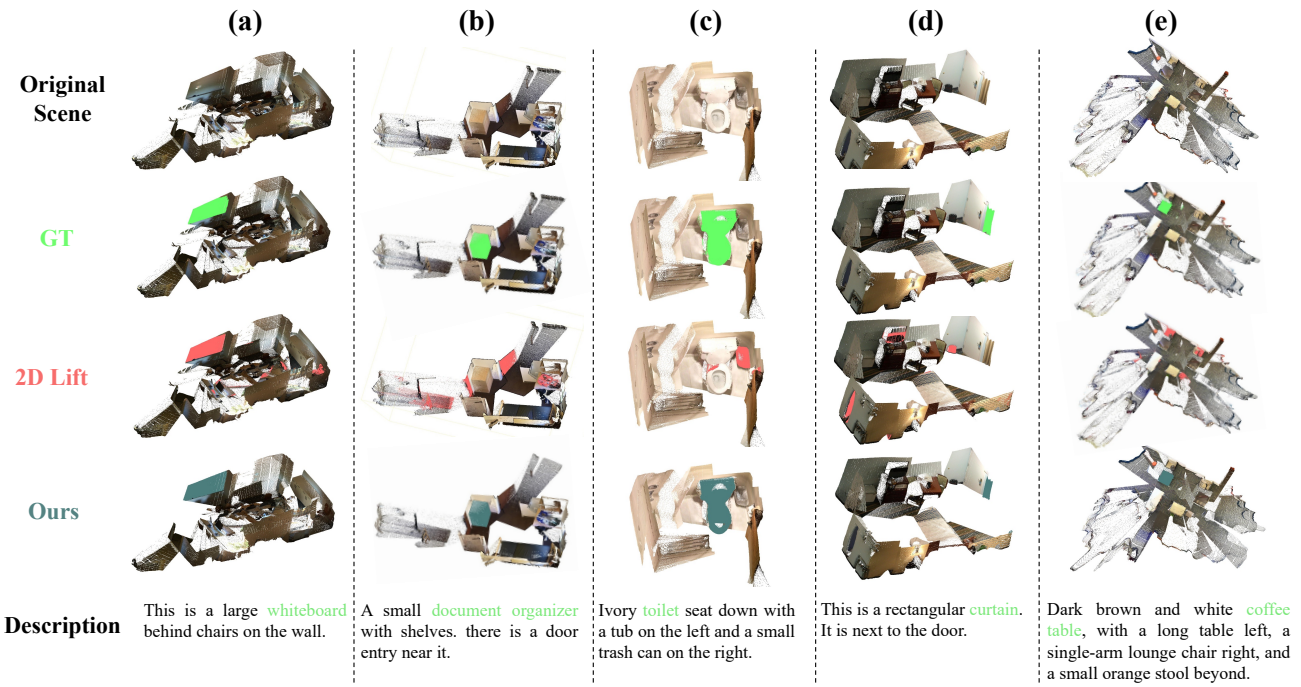


Figure 4. Qualitative comparison on the MVRefer benchmark.

Table 5. Ablation on MVGGT fusion stage.

Fusion Stage	Easy ↑		Hard ↑		Overall ↑	
	$mIoU_{global}$	$mIoU_{view}$	$mIoU_{global}$	$mIoU_{view}$	$mIoU_{global}$	$mIoU_{view}$
Early	45.7	65.9	21.6	63.4	36.1	64.9
Middle	47.5	65.7	22.5	62.1	37.5	64.3
Late	**50.1**	**70.6**	**24.4**	**67.3**	**39.9**	**69.3**

Multimodal Branch Depth. Table 6 evaluates branch depth. A 6-layer configuration lacks alignment capacity, while 12 layers yield peak performance. Expanding to 16 layers causes a sharp decline due to overfitting and unstable attention under sparse supervision. These results suggest that moderate depth optimally balances expressiveness and regularity, ensuring language effectively guides aggregation without overwhelming the geometric signal.

Table 6. Ablation on layer number of multimodal branch.

Layers	Easy ↑		Hard ↑		Overall ↑	
	$mIoU_{global}$	$mIoU_{view}$	$mIoU_{global}$	$mIoU_{view}$	$mIoU_{global}$	$mIoU_{view}$
6	47.1	66.7	22.5	64.1	37.3	65.7
12	**50.1**	**70.6**	**24.4**	**67.3**	**39.9**	**69.3**
16	35.6	26.2	10.1	14.2	25.4	21.4

5.4. Qualitative Analysis

Fig. 3 demonstrates MVGGT's robustness under sparse and noisy views. Unlike 2D-lifting baselines that often drift or collapse under occlusion and depth ambiguity, MVGGT maintains stable, target-aligned segmentations. Specifically, it leverages geometric-linguistic fusion to resolve planar ambiguities (a) and semantic distraction (c). PVSO-balanced supervision further enables precise localization in clutter (b), while the architecture ensures fine-grained discrimination (d) and cross-view consistency despite partial visibility (e).

6. Conclusion

We introduce MV-3DRES, a sparse-view 3D grounding setting reflecting real-world agent conditions. To overcome the failure of conventional two-stage pipelines under sparsity, we propose MVGGT, a dual-branch architecture that integrates linguistic cues directly into geometric reasoning. We further address the Foreground Gradient Dilution challenge via PVSO, enabling stable optimization. Together with the MVRefer benchmark, our framework provides a practical path for robust multimodal 3D grounding and embodied perception.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2025YFE0113500), National Science Fund for Distinguished Young Scholars (No.62525605) and the National Natural Science Foundation of China (No. U25B2066, No. U22B2051, No. 62302411).

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 2, 3

[2] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1050–1060, 2025. 3

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 2, 3, 4, 6

[4] Wenxin Chen, Mengxue Qu, Weitai Kang, Yan Yan, Yao Zhao, and Yunchao Wei. 3drest: A strong baseline for semi-supervised 3d referring expression segmentation. *arXiv preprint arXiv:2504.12599*, 2025. 3

[5] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. 3

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 4

[7] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4): 1, 2017. 2

[8] 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. 3, 7

[9] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025. 3

[10] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *2025 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2025. 3

[11] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3r-sfm: Towards feed-forward structure-from-motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16774–16784, 2025. 3

[12] Jiading Fang, Xiangshan Tan, Shengjie Lin, Igor Vasiljevic, Vitor Guizilini, Hongyuan Mei, Rares Ambrus, Gregory Shakhnarovich, and Matthew R Walter. Transcrib3d: 3d referring expression resolution through large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9737–9744. IEEE, 2024. 3

[13] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. *Advances in neural information processing systems*, 37:57207–57239, 2024. 3

[14] 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 *ICCV*, 2021. 3

[15] Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5185–5193, 2021. 3

[16] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4313–4322, 2022. 3

[17] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15372–15383, 2023. 3

[18] 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, 2021. 3

[19] Shuting He and Henghui Ding. Refmask3d: Language-guided transformer for 3d referring segmentation. In *ACM MM*, 2024. 3

[20] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *ECCV*, 2024. 3, 7

[21] JiaKui Hu, Shanshan Zhao, Qing-Guo Chen, Xuerui Qiu, Jialun Liu, Zhao Xu, Weihua Luo, Kaifu Zhang, and Yanye Lu. Omni-view: Unlocking how generation facilitates understanding in unified 3d model based on multiview images. *arXiv preprint arXiv:2511.07222*, 2025. 3

[22] Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation. *arXiv preprint arXiv:2503.18135*, 2025.

[23] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In *International Conference on 3D Vision 2025*, 2025. 3, 7

[24] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, 2021. 2, 3, 7

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

[26] Zhuoxu Huang, Mingqi Gao, and Jungong Han. Point linguist model: Segment any object via bridged large 3d-language model. *arXiv preprint arXiv:2509.07825*, 2025. 3

[27] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2989–2998, 2023. 3

[28] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3

[29] Yongmin Kim, Chenhui Chu, and Sadao Kurohashi. Flexible visual grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 285–299, 2022. 3

[30] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8500–8509, 2022. 3

[31] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 3

[32] Hongyang Li, Jinyuan Qu, and Lei Zhang. Ovseg3r: Learn open-vocabulary instance segmentation from 2d via 3d reconstruction. *arXiv preprint arXiv:2509.23541*, 2025. 3

[33] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Transformer-empowered invariant grounding for video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[34] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024. 3

[35] Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. *arXiv preprint arXiv:2501.14607*, 2025. 6

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

[37] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2(1):16, 2024. 3

[38] Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6032–6041, 2021. 3

[39] Xuexun Liu, Xiaoxu Xu, Jinlong Li, Qiudan Zhang, Xu Wang, Nicu Sebe, and Lin Ma. Less: Label-efficient and single-stage referring 3d segmentation. *arXiv:2410.13294*, 2024. 3, 6, 7

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6

[41] Yihang Liu, Changli Wu, Xiaoshuai Sun, Jiayi Ji, Yiwei Ma, Gen Luo, Liujuan Cao, and Rongrong Ji. Weakly-supervised 3d referring expression segmentation. 2024. 3

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[43] 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 *CVPR*, 2022. 3

[44] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 3

[45] Zador Pataki, Paul-Edouard Sarlin, Johannes L Schönberger, and Marc Pollefeys. Mp-sfm: Monocular surface priors for robust structure-from-motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21891–21901, 2025. 3

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

[47] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *AAAI*, 2024. 3

[48] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025. 3

[49] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 3

[50] Grounded Transformer. Iggt: Instance-grounded geometry trans-former for semantic 3d reconstruction. 3

[51] Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster vggt with block-sparse global attention. *arXiv preprint arXiv:2509.07120*, 2025. 3

[52] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3

[53] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 3

[54] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 3

[55] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 20697–20709, 2024. 3

[56] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Pi3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 3, 6

[57] Yi Ru Wang, Yuchi Zhao, Haoping Xu, Saggi Eppel, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Mvtrans: Multi-view perception of transparent objects. *arXiv preprint arXiv:2302.11683*, 2023. 3

[58] WinT3R WinT3R. Wint3r: Window-based streaming reconstruction with camera token pool. 3

[59] Changli Wu, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, Rongrong Ji, et al. Rgsan: Rule-guided spatial awareness network for end-to-end 3d referring expression segmentation. *Advances in Neural Information Processing Systems*, 37:110972–110999, 2024. 2, 3, 7

[60] 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 *ACM MM*, 2024.

[61] 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 *AAAI*, 2024. 2, 3, 7

[62] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024. 3

[63] Qi Xu, Dongxu Wei, Lingzhe Zhao, Wenpu Li, Zhangchi Huang, Shunping Ji, and Peidong Liu. Siu3r: Simultaneous scene understanding and 3d reconstruction beyond feature alignment. *arXiv preprint arXiv:2507.02705*, 2025. 3

[64] 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 *ICCV*, 2021. 3

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 5

[66] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*, 2023. 2

[67] Zhi Zhang, Helen Yannakoudakis, Xiantong Zhen, and Ekaterina Shutova. Ck-transformer: Commonsense knowledge enhanced transformers for referring expression comprehension. *arXiv preprint arXiv:2302.09027*, 2023. 3

[68] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, 2021. 2