ANCHORSPLAT: FEED-FORWARD 3D REFERRING SEGMENTATION IN GAUSSIAN SCENES

Anonymous authors
Paper under double-blind review

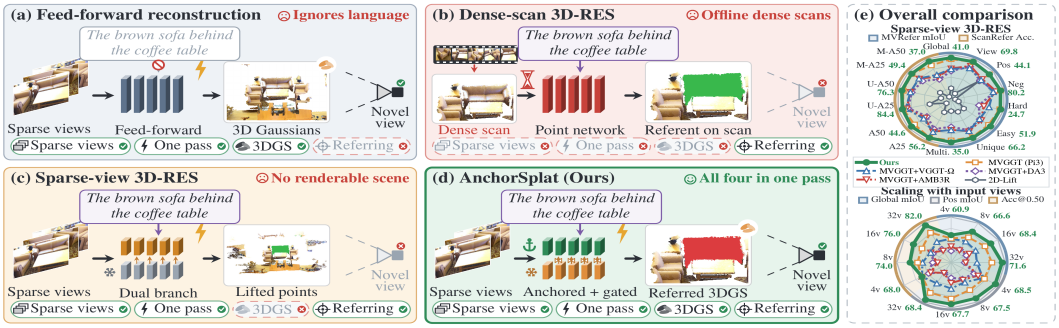


Figure 1: Four task formulations for one ScanRefer expression on a ScanNet scene. AnchorSplat grounds the expression from sparse uncalibrated views in a renderable Gaussian scene. Green marks the target and red marks the prediction.

ABSTRACT

Grounding a referring expression in sparse, uncalibrated views requires both semantic discrimination and 3D consistency. Feed-forward reconstruction models supply cameras, depth, and renderable Gaussians, yet substituting them for the backbone of a multiview referring model consistently lowers accuracy. We trace this failure to two mechanisms. Reconstruction pretraining progressively attenuates the appearance cues that referring expressions rely on, and pixel-aligned Gaussian rendering allows each primitive to copy the label of the pixel that created it. We present AnchorSplat, which predicts view masks and a referred Gaussian scene in a single forward pass. A frozen semantic encoder anchors the referring state, and zero-initialized gates admit reconstruction features while preserving its initial referring function. A frame-invariant correspondence graph transports language-conditioned states across views, and an opacity-gated readout labels each Gaussian under owner-excluded rendering, which evaluates a primitive only in views that did not create it. On the 9,508 test samples of MVRefer, AnchorSplat improves global mIoU from 39.9 to 41.0 and Acc@0.50 from 41.5 to 44.6 over MVGGT, whereas direct substitutions with VGGT-Ω, DA3, and AMB3R reach at most 32.8 global mIoU.

1 INTRODUCTION

Language-guided agents must localize referred objects in 3D from sparse, uncalibrated views (Wang et al., 2024b). Most methods require registered scans or dense posed images (Chen et al., 2020; Achlioptas et al., 2020; Huang et al., 2021; Wu et al., 2024a); language-embedded radiance fields and Gaussian scenes further require per-scene optimization (Kerr et al., 2023; Qin et al., 2024; He et al., 2025). MVGGT removes the scan requirement but produces view masks and lifted points (Wu et al., 2026). Feed-forward Gaussian reconstruction could instead provide an editable referent that can be verified from novel views (Figure 1) (Jiang et al., 2025; Lin et al., 2026).

Simply replacing MVGGT's Pi3 backbone (Wang et al., 2026b) fails: VGGT-Ω, Depth Anything 3 (DA3), and AMB3R lower global mIoU from 39.9 to 32.5, 30.9, and 32.8 (Figure 2a) (Wang et al.,

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.

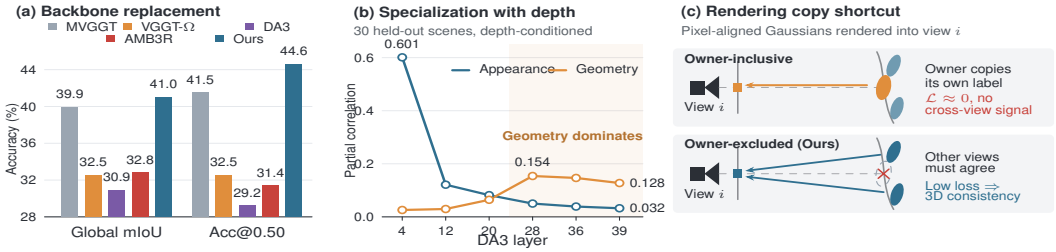


Figure 2: Two reconstruction obstacles: (a) weaker direct substitutions on MVRefer; (b) appearance erosion and geometric specialization across DA3 layers in 30 scenes; (c) the source-view label-copy shortcut.

2026a; Lin et al., 2026; Wang & Agapito, 2026). Reconstruction pretraining favors viewpoint and texture invariance, whereas expressions often distinguish color or material. In DA3, appearance evidence falls by 94.6% from layer 4 to 39 while geometric evidence peaks at 5.9 times its shallow value (Figure 2b). We call this *semantic erosion under geometric specialization*.

Rendering poses another problem. A pixel-aligned Gaussian can dominate its source pixel, allowing a source-view rendering loss to be minimized by copying that pixel's label without cross-view agreement (Figure 2c). MVGGT supervises views separately; feed-forward Gaussian reconstruction has no referring labels; and semantic Gaussian fields typically require per-scene optimization (Wu et al., 2026; Jiang et al., 2025; Lin et al., 2026; Qin et al., 2024; He et al., 2025). Reconstruction should instead route semantic content across its feature hierarchy, predicted 3D correspondences, and pixel-to-primitive readout.

AnchorSplat combines frozen DINOv2 semantics with frozen DA3 geometry (Oquab et al., 2024; Lin et al., 2026). Zero-initialized gates preserve the initial referring function, a frame-invariant graph transports language-conditioned states across views, and owner-excluded rendering supervises an opacity-gated Gaussian readout. Footprint refinement improves novel-view rendering without changing segmentation. The single-pass model returns view masks and a referred Gaussian scene (Figure 3), improving 17 of 20 MVRefer mIoU entries and all nine ScanRefer entries over MVGGT.

Our contributions are:

- We identify semantic erosion and the owner-copy shortcut as two obstacles to using reconstruction models for referring segmentation.
- We propose AnchorSplat, which grounds a free-form expression in a renderable Gaussian scene from sparse uncalibrated views in one forward pass.
- A semantic-anchored ladder preserves the initial referring function, while frame-invariant correspondences transport its states across views.
- Owner-excluded rendering converts a low rendering loss into cross-view agreement for an opacity-gated Gaussian readout.

2 Related Work

3D referring segmentation. ScanRefer and ReferIt3D assume registered scans (Chen et al., 2020; Achlioptas et al., 2020; Dai et al., 2017). Later graph, superpoint, spatial, and language-model methods retain this assumption (Huang et al., 2021; Wu et al., 2024b;a; Liu et al., 2024; Huang et al., 2025; He et al., 2024; Deng et al., 2025). MVGGT grounds sparse RGB views with frozen Pi3 features (Wu et al., 2026; Wang et al., 2026b); AnchorSplat additionally grounds the referent in Gaussian primitives.

Feed-forward geometry and Gaussians. DUSt3R and MASt3R predict pointmaps and matches (Wang et al., 2024a; Leroy et al., 2024). VGGT, VGGT-Ω, Pi3, DA3, and AMB3R extend reconstruction to many views (Wang et al., 2025; 2026a;b; Lin et al., 2026; Wang & Agapito, 2026). Differentiable 3D Gaussians (Kerbl et al., 2023; Ye et al., 2025b) can be predicted from posed views

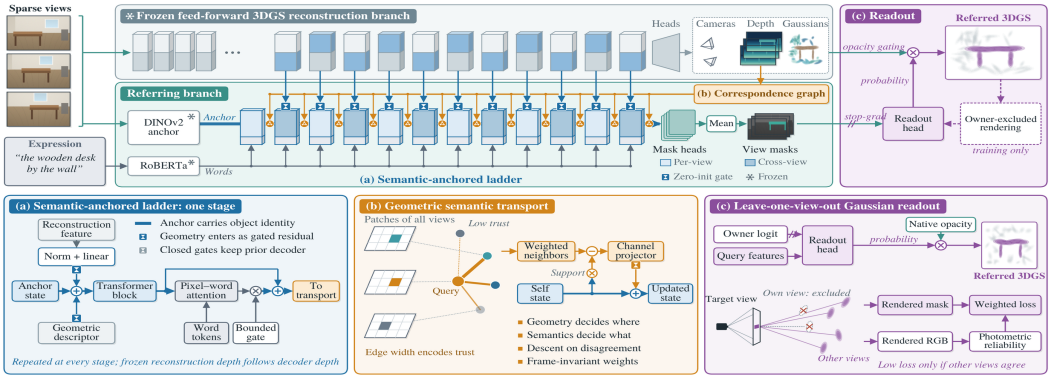


Figure 3: AnchorSplat anchors DINOv2 semantics, admits DA3 geometry through gates, transports states across views, and labels Gaussians through an opacity-gated readout.

by pixelSplat and MVSplat, or unposed views by NoPoSplat, AnySplat, DA3, and ZipSplat (Charatan et al., 2024; Chen et al., 2024; Ye et al., 2025a; Jiang et al., 2025; Lin et al., 2026; Veicht et al., 2026). Their geometry and photometry objectives do not preserve appearance cues needed for referring.

Language and semantic Gaussian fields. LERF, LangSplat, Feature 3DGS, Gaussian Grouping, and ReferSplat attach language or semantic content to optimized scenes (Kerr et al., 2023; Qin et al., 2024; Zhou et al., 2024; Ye et al., 2024; He et al., 2025). LSM predicts open-vocabulary fields from unposed images (Fan et al., 2024), but does not address free-form referring. AnchorSplat grounds such expressions in a feed-forward Gaussian scene.

3 Problem Setting

Let $\mathcal{I} = \{I_i\}_{i=1}^{N}$ be sparse RGB views of a static scene with unknown intrinsics and poses, and let q be a free-form referring expression. We learn a single network $f(\mathcal{I}, q) = (\mathcal{C}, \mathcal{D}, \mathcal{M}, \mathcal{G}, \mathbf{s})$ predicts cameras $\mathcal{C}$, depths $\mathcal{D}$, view logits $\mathcal{M}$, a Gaussian scene $\mathcal{G}$, and primitive referring probabilities $\mathbf{s}$. Primitive j has center $\boldsymbol{\mu}_j$, scale $\boldsymbol{\sigma}_j$, rotation $\mathbf{r}_j$, opacity o_j, and appearance $\mathbf{h}_j$. The referred scene is

$$\mathcal{G}_q = \{(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \mathbf{r}_j, \mathbf{h}_j,\ o_j s_j)\}_{j=1}^{J}. \tag{1}$$

Because only opacity depends on q, the referred scene shares its geometry with the full scene and remains renderable from any viewpoint. We evaluate view masks and lifted points under the MVRefer protocol (Wu et al., 2026) and additionally evaluate full-scene and referred-scene rendering.

4 Method

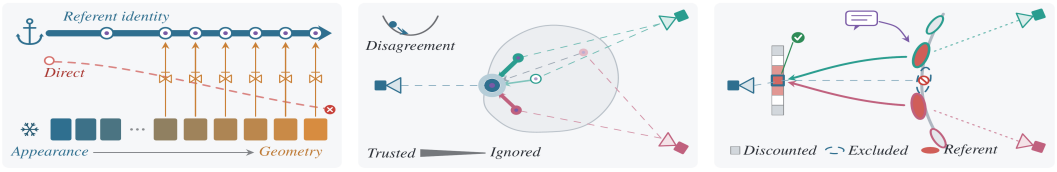
AnchorSplat preserves referent identity across three transitions, from shallow to deep reconstruction features, across views, and from pixels to Gaussian primitives (Figure 4). Sections 4.2 to 4.4 address them in turn.

4.1 Gaussian Scene and Semantic Anchor

Frozen DA3 predicts cameras, depths, confidences, and pixel-aligned Gaussians (Lin et al., 2026). With intrinsics K_i and homogeneous world-to-camera extrinsics E_i, pixel $\mathbf{u} = (u, v, 1)^\top$ of view i lifts to

$$\begin{bmatrix} \mathbf{p}_i(\mathbf{u}) \\ 1 \end{bmatrix} = E_i^{-1} \begin{bmatrix} D_i(\mathbf{u}) K_i^{-1} \mathbf{u} \\ 1 \end{bmatrix}. \tag{2}$$

The 40 alternating-attention blocks of DA3 provide view-local and cross-view reconstruction features. The semantic anchor $S \in \mathbb{R}^{N \times P \times d}$ consists of frozen DINOv2 patch tokens with registers (Oquab et al., 2024; Darcet et al., 2024), which retain the feature basis of the MVGGT referring decoder. Frozen RoBERTa encodes q into word tokens $T \in \mathbb{R}^{n_w \times d}$ (Liu et al., 2019), and learned register tokens are added to each view.



(a) Semantic ladder (b) Geometric transport (c) Held-out rendering

Figure 4: Three roles of geometry: (a) gated features, (b) cross-view transport, and (c) owner-excluded Gaussian rendering.

4.2 Semantic-Anchored Geometry Ladder

The decoder has twelve stages that alternate view-local and scene-wide attention. Stage 1 reads S, stages 2 to 11 alternately read the local and global halves of DA3 blocks 28 to 37, and stage 12 reads the global half of block 39, preserving the native block order. With patch state Z^ℓ and assigned reconstruction feature X^ℓ, the stage input is

$$\widetilde{Z}^\ell = Z^{\ell-1} + \boldsymbol{\gamma}_\ell \odot \phi_\ell\big(X^\ell\big) + \boldsymbol{\beta}_\ell + W_G\,\mathbf{g}, \tag{3}$$

where ϕ_ℓ is a normalized linear projection, $(\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell)$ is a zero-initialized per-channel gate, and the zero-initialized W_G injects the geometric descriptor $\mathbf{g}$ of Section 4.3. With a transformer block B_ℓ and a pixel-word attention F_ℓ, the language update is

$$U^\ell = B_\ell(\widetilde{Z}^\ell), \qquad L^\ell = F_\ell(U^\ell, T), \qquad \widehat{Z}^\ell = U^\ell + \tanh\big(R(L^\ell)\big) \odot L^\ell, \tag{4}$$

where $\tanh(R(\cdot))$ bounds the language contribution at each stage.

Proposition 1 (Semantic preservation)**.** *Let θ_g collect the gates $(\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell)_{\ell\geq 2}$, the descriptor map W_G, and the transport parameters $(W_Q, \mathbf{b}_Q)$ of Eq. equation 6. At $\theta_g = 0$, every view logit of AnchorSplat equals the logit of the transferred referring decoder driven by S alone. If each stage is Γ_ℓ-Lipschitz, the deviation of the final state from this reference is bounded by the sum of the injection norms $\|\boldsymbol{\gamma}_\ell \odot \phi_\ell(X^\ell) + \boldsymbol{\beta}_\ell + W_G\mathbf{g}\|$ and the transport increments, each weighted by the product of the downstream constants Γ_ℓ.*

Proposition 1 provides an exact semantic starting point and bounds the drift introduced by geometry. Direct substitution lacks this reference because its decoder faces a new feature basis from the first update. Since a zero gate blocks the initial gradient to ϕ_ℓ, each ϕ_ℓ is initialized by aligning reconstruction tokens with the anchor basis.

4.3 Geometric Semantic Transport

Each pixel carries a nine-dimensional descriptor of position, scale, opacity, depth, and confidence, normalized by confidence-weighted centering and scaling to remove the arbitrary origin and scale of the reconstruction frame. Patch pooling yields a descriptor $\mathbf{g}_a$, a normalized position $\bar{\mathbf{p}}_a$, and a confidence c_a. Each patch a connects to its four nearest patches $\mathcal{N}(a)$ in other views, with $d_{ab} = \|\bar{\mathbf{p}}_a - \bar{\mathbf{p}}_b\|$ and edge weights

$$w_{ab} = \underbrace{\frac{c_b\, e^{-\tau d_{ab}}}{\sum_{k\in\mathcal{N}(a)} c_k\, e^{-\tau d_{ak}}}}_{\pi_{ab}} \underbrace{e^{-\kappa \min_k d_{ak}}}_{\rho_a}, \qquad \tau = 20,\ \kappa = 10, \tag{5}$$

where π_{ab} distributes weight by proximity and confidence and ρ_a attenuates transfer when no close match exists, giving the support mass $m_a = \sum_b w_{ab} = \rho_a$. After each language update, the state moves along the graph,

$$Z_a^\ell = \widehat{Z}_a^\ell + W_Q\Big(\sum_{b\in\mathcal{N}(a)} w_{ab}\widehat{Z}_b^\ell - m_a\widehat{Z}_a^\ell\Big) + m_a\mathbf{b}_Q, \tag{6}$$

with W_Q and $\mathbf{b}_Q$ initialized at zero. The bracket is the negative local gradient of $\mathcal{E}_a(Z) = \frac{1}{2}\sum_{b\in\mathcal{N}(a)} w_{ab}\|Z_a - Z_b\|_2^2$.

Proposition 2 (Bounded, frame-invariant transport). *(i)* $0 < m_a \leq 1$. *(ii) The bracket in Eq. equation 6 equals* $-\nabla_{Z_a}\mathcal{E}_a$. *(iii) If* $W_Q = \eta\Pi$ *with* $\Pi \succeq 0$ *and* $0 < \eta\lambda_{\max}(\Pi) < 2$, *the linear part of the update strictly decreases* $\mathcal{E}_a$ *with the neighbors held fixed whenever* $\Pi\nabla_{Z_a}\mathcal{E}_a \neq 0$. *For* $\Pi = I$ *and* $0 \leq \eta \leq 1$, *the new state is a convex combination of* $\widehat{Z}_a$ *and its neighbors. (iv) Every* w_{ab} *is invariant to a similarity transform of the reconstruction frame.*

With $W = [w_{ab}]$, the bracket equals $-(\mathrm{diag}(\mathbf{m}) - W)\widehat{Z}$, where $\mathrm{diag}(\mathbf{m}) - W$ is a directed graph Laplacian (Zhou et al., 2003). Transport is thus a learned diffusion step on the correspondence graph, in which geometry selects where referring content moves and W_Q selects which channels move.

4.4 Leave-One-View-Out Gaussian Readout

Primitive j is predicted from its owner pixel $\mathbf{u}_j$ in view i_j, and its referring logit is

$$z_j = \mathrm{sg}\big[M_{i_j}(\mathbf{u}_j)\big] + H(\mathbf{v}_j \odot \mathrm{Attn}(\mathbf{v}_j, T) + W_C\,\mathbf{g}_j)\,, \qquad s_j = \mathrm{sigmoid}(z_j), \tag{7}$$

where $\mathbf{v}_j$ and $\mathbf{g}_j$ are the projected state and descriptor at the owner patch, and sg stops gradients. The readout starts from the view logit of its owner, and detached inputs confine rendering gradients to H.

Owner-excluded rendering. Rendering labels into their source views would admit the owner-copy shortcut. We therefore render view i without the primitives $\mathcal{O}_i$ it owns,

$$\widehat{P}_i^{\setminus i}(\mathbf{u}) = \frac{\sum_{j\notin\mathcal{O}_i}\omega_j(\mathbf{u})\,s_j}{A_i^{\setminus i}(\mathbf{u})}, \qquad A_i^{\setminus i}(\mathbf{u}) = \sum_{j\notin\mathcal{O}_i}\omega_j(\mathbf{u}), \tag{8}$$

where $\omega_j(\mathbf{u})$ is the compositing weight. Primitives owned by view i thus receive no gradient from its supervision (Figure 4c).

Proposition 3 (Cross-view agreement). *Let* $\bar{\omega}_j(\mathbf{u}) = \omega_j(\mathbf{u})/A_i^{\setminus i}(\mathbf{u})$ *be the normalized weight of primitive* $j \notin \mathcal{O}_i$ *at pixel* $\mathbf{u}$ *of view* i. *If* $y_i(\mathbf{u}) = 1$ *and* $\widehat{P}_i^{\setminus i}(\mathbf{u}) \geq 1 - \epsilon$, *then* $\bar{\omega}_j(\mathbf{u})(1 - s_j) \leq \epsilon$ *for every such* j. *If* $y_i(\mathbf{u}) = 0$ *and* $\widehat{P}_i^{\setminus i}(\mathbf{u}) \leq \epsilon$, *then* $\bar{\omega}_j(\mathbf{u})\,s_j \leq \epsilon$.

Proposition 3 links low rendering loss to cross-view label agreement. We weight supervision by $\Omega_i(\mathbf{u}) = A_i^{\setminus i}(\mathbf{u})\exp(-8\|\widehat{I}_i^{\setminus i}(\mathbf{u}) - I_i(\mathbf{u})\|_1/3)$ to discount unreliable coverage.

4.5 Render-Only Footprint Refinement

Sparse-view pose and depth errors leave holes in novel-view renders (Charatan et al., 2024; Chen et al., 2024; Ye et al., 2025a). A lightweight head Ψ predicts bounded residuals from frozen reconstruction features and source appearance, and only the footprint residual $\boldsymbol{\delta}_j$ is retained,

$$\boldsymbol{\sigma}_j^{\mathrm{r}} = \boldsymbol{\sigma}_j \odot \exp(0.5\tanh\boldsymbol{\delta}_j)\,, \qquad \big(\boldsymbol{\mu}_j^{\mathrm{r}}, \mathbf{r}_j^{\mathrm{r}}, \mathbf{h}_j^{\mathrm{r}}, o_j^{\mathrm{r}}\big) = (\boldsymbol{\mu}_j, \mathbf{r}_j, \mathbf{h}_j, o_j)\,. \tag{9}$$

Trained on held-out views of training scenes, Ψ changes only primitive footprints, so $\mathcal{G}_q$ and all segmentation outputs remain unchanged.

4.6 Training and Inference

Mask decoders at stages 4, 8, 11, and 12 use binary cross-entropy and view-balanced Dice losses, and the readout uses the reliability-weighted rendering loss and a class-balanced owner loss,

$$\mathcal{L} = \mathcal{L}_{\mathrm{mask}}^{12} + \lambda_{\mathrm{aux}}\sum_{k\in\{4,8,11\}}\mathcal{L}_{\mathrm{mask}}^{k} + \lambda_{\mathrm{render}}\mathcal{L}_{\mathrm{render}} + \lambda_{\mathrm{owner}}\mathcal{L}_{\mathrm{owner}}. \tag{10}$$

We first train the ladder with transport inactive, then calibrate the descriptor and transport maps with other parameters frozen, and finally fit Ψ. At inference, averaging the four stage logits yields a product-of-experts decision (Hinton, 2002). Proofs appear in the supplementary material.

Table 1: Comparison on MVRefer (mIoU, %). Pos and Neg denote target-visible and target-absent views.

	Hard (∼40%)				Easy (∼60%)				Unique (∼19%)				Multiple (∼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 (Wu et al., 2026)	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., 2026)	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., 2026)	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
MVGGT+VGGT-Ω (Wang et al., 2026a)	19.7	59.0	29.0	68.2	41.1	60.4	45.0	69.9	57.0	72.0	59.0	79.7	27.0	57.2	34.0	66.8	32.5	59.9	38.6	69.2
MVGGT+DA3 (Lin et al., 2026)	15.3	59.7	24.4	70.3	41.3	63.5	46.7	74.0	55.8	75.8	58.4	84.7	24.9	58.7	32.8	69.6	30.9	62.0	37.8	72.6
MVGGT+AMB3R (Wang & Agapito, 2026)	17.4	61.1	26.7	71.5	43.0	64.6	47.7	74.8	57.4	76.9	59.2	86.1	26.9	59.9	34.5	70.5	32.8	63.2	39.3	73.5
AnchorSplat	**24.7**	67.1	31.0	77.6	**51.9**	**71.7**	**52.9**	**81.9**	**66.2**	**83.2**	**65.1**	**90.8**	**35.0**	**66.6**	**39.1**	**77.6**	**41.0**	**69.8**	**44.1**	**80.2**

Table 2: Comparison under the ScanRefer protocol (%). Ranks are computed within each input setting.

	Unique (∼19%)			Multiple (∼81%)			Overall		
Method	A25	A50	mIoU	A25	A50	mIoU	A25	A50	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)	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	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., 2024)	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	53.2	29.9	33.7
3D-LLaVA (Deng et al., 2025)	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	43.3
MV-3DRES with sparse RGB views only									
Two-stage (Wu et al., 2026)	43.8	20.9	27.5	25.1	12.2	16.4	28.7	13.9	18.5
2D-Lift (Wu et al., 2026)	51.6	26.0	31.5	21.5	5.6	14.5	27.3	9.6	17.8
MVGGT (Wu et al., 2026)	83.6	74.5	65.2	49.2	33.5	33.8	55.9	41.5	39.9
MVGGT+VGGT-Ω (Wang et al., 2026a)	80.4	65.2	57.0	42.3	25.1	27.0	49.3	32.5	32.5
MVGGT+DA3 (Lin et al., 2026)	77.6	63.5	55.8	38.7	21.0	24.9	46.3	29.2	30.9
MVGGT+AMB3R (Wang & Agapito, 2026)	79.1	65.3	57.4	42.1	23.3	26.9	49.3	31.4	32.8
AnchorSplat	**84.4**	**76.3**	**66.2**	**49.4**	**37.0**	**35.0**	**56.2**	**44.6**	**41.0**

5 Experiments

5.1 Setup

Benchmark and metrics. MVRefer (Wu et al., 2026) pairs ScanRefer expressions with ScanNet sequences (Chen et al., 2020; Dai et al., 2017), and its test split contains 9,508 samples with eight 518×336 views. Hard targets occupy less than 5% of every visible view, and Unique scenes contain no same-class distractor. Global mIoU pools all views of a sample and equals lifted-point IoU. View, Pos, and Neg mIoU average over all, target-visible, and target-absent views, and Acc@0.25 and Acc@0.50 threshold the global IoU. Colors mark the **first**, second, and third results within each setting.

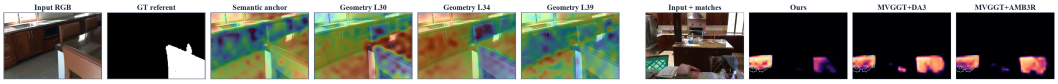
Rendering protocol. Held-out ScanNet frames serve as rendering targets. Methods receive eight uncalibrated views, except the two-view NoPoSplat release, and a single similarity transform aligns predicted and target cameras without per-scene optimization. PSNR, SSIM, and LPIPS are reported on full frames and referent regions.

5.2 Comparison with State-of-the-Art Methods

AnchorSplat achieves the best overall results on MVRefer (Table 1), improving 17 of 20 MVGGT entries and raising global mIoU from 39.9 to 41.0, whereas all direct substitutions fall 7.1 to 9.0 points below MVGGT. Under the ScanRefer protocol (Table 2), AnchorSplat improves all nine entries and raises Acc@0.50 from 41.5 to 44.6, within 0.3 points of scan-based RG-SAN while using only sparse RGB views.

5.3 Semantic Erosion and View Scaling

Semantic erosion. The DA3 probe shows appearance correlation falling from 0.601 to 0.032 and geometry correlation rising from 0.026 to 0.128 across layers 4–39 (Figure 2b). The anchored decoder reduces distractor activation (Figure 5).



(a) Appearance-word attention across semantic and geometric stages.

(b) Final-stage referring probabilities of AnchorSplat and direct substitutions.

Figure 5: Appearance anchoring and referent discrimination.

View scaling and cost. From 4 to 32 views, AnchorSplat gains 10.66 global mIoU points compared with 6.95 for MVGGT, whereas direct substitutions stagnate or degrade (Table 3a). Geometric transport therefore turns additional views into referring evidence. The semantic, language, and Gaussian modules add only 0.067 s to an eight-view pass (Table 3b).

Table 3: View scaling on 50 trajectories and eight-view runtime profile. Colors mark first, second, and third per metric among methods at the same view count in (a), and among methods in (b); ties share rank. Higher is better for accuracy and lower is better for time and memory. The constant Other column and single available Gauss. entry are unranked. Times are medians of 24 warmed-up runs.

(a) Nested view-count diagnostic

Method	Views	Global	View	Pos	Acc@0.50	Mask wall (s)	Mask peak (GB)
MVGGT (Pi3)	4	58.68	**73.64**	61.60	**68.00**	0.15	**5.93**
	8	62.01	73.29	61.26	68.00	0.26	**6.12**
	16	64.60	77.28	64.09	68.00	0.62	**6.58**
	32	65.63	77.06	63.36	74.00	1.55	**7.49**
MVGGT+VGGT-Ω	4	58.49	71.29	60.84	**68.00**	**0.14**	6.43
	8	58.93	70.18	55.70	64.00	**0.23**	6.63
	16	58.48	69.58	57.68	60.00	**0.52**	7.10
	32	56.99	68.32	56.41	64.00	**1.25**	8.06
MVGGT+DA3	4	50.08	62.69	55.24	62.00	0.19	7.91
	8	52.32	68.24	53.17	56.00	0.36	8.71
	16	52.11	67.20	50.85	58.00	0.80	10.35
	32	49.94	65.71	48.30	56.00	1.91	13.65
MVGGT+AMB3R	4	53.18	67.38	57.82	60.00	0.46	6.51
	8	51.80	66.51	53.75	52.00	0.83	7.64
	16	54.15	70.24	54.20	60.00	1.85	8.48
	32	50.59	66.42	48.75	54.00	4.40	10.81
AnchorSplat	4	**60.90**	72.85	**68.50**	**68.00**	0.32	9.68
	8	**66.56**	**77.48**	**67.50**	**74.00**	0.59	11.38
	16	**68.43**	**79.80**	**67.69**	**76.00**	1.27	12.86
	32	**71.56**	**81.34**	**68.44**	**82.00**	2.86	16.53

(b) Complete eight-view forward pass

Method	Geom.	Sem./text	Fuse/trans.	Gauss.	Other	CUDA	Wall	Peak
MVGGT (Pi3)	**0.072**	**0.002**	0.213	N/A	0.002	0.289	0.289	**6.11**
MVGGT+VGGT-Ω	0.177	0.005	**0.066**	N/A	0.002	**0.251**	**0.251**	7.13
MVGGT+DA3	0.360	**0.002**	0.093	N/A	0.002	0.457	0.457	8.81
MVGGT+AMB3R	0.727	0.005	0.093	N/A	0.002	0.826	0.826	7.63
AnchorSplat	0.433	0.052	0.094	0.015	0.002	0.596	0.594	11.46

5.4 Referred Gaussian Scene

Footprint refinement raises the full-frame PSNR of the DA3 scene from 13.90 to 14.26 dB and attains the best SSIM and LPIPS in both full frames and referent regions (Table 4). Refinements that alter appearance, fuse features freely, feed renders back, or add surfels all lower full-frame quality, confirming footprints as the appropriate degree of freedom. Rendering $\mathcal{G}_q$ alone removes the background by design, yet its object-region PSNR of 13.95 dB exceeds that of AnySplat and NoPoSplat rendering complete scenes, indicating that opacity gating isolates the referent faithfully (Figures 6 and 7).

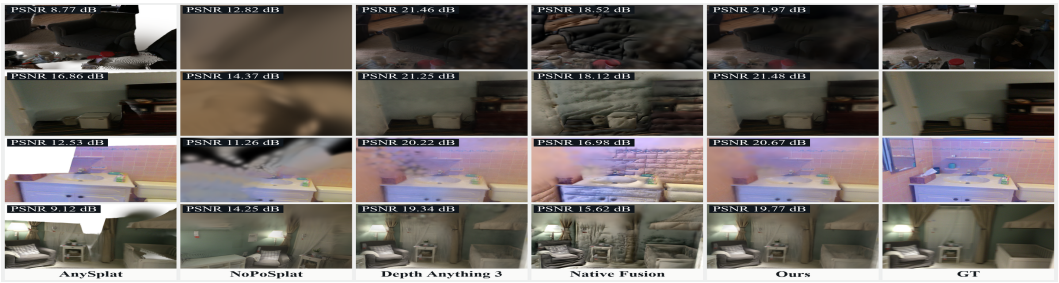


Figure 6: Full-frame renders of held-out views. Insets report PSNR in dB.

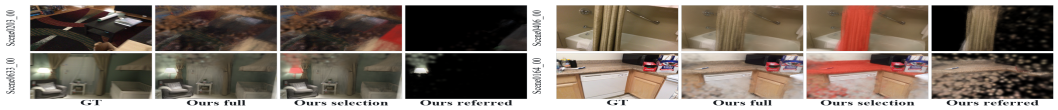


Figure 7: Referred Gaussian selection and rendering in four target-visible views. Red encodes the referring probability.

Table 4: Novel-view synthesis on 111 ScanNet views. Colors rank eight-view complete scenes; † uses two views. The last row renders only $\mathcal{G}_q$.

	Full frame			Object region		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AnySplat (Jiang et al., 2025)	8.75	0.395	0.684	10.08	0.244	0.058
NoPoSplat† (Ye et al., 2025a)	10.42	0.381	0.818	12.74	0.332	0.063
Direct MVGGT substitutions			N/A			
Native DA3 scene (Lin et al., 2026)	13.90	0.458	0.651	16.91	0.385	0.044
AnchorSplat	**14.26**	**0.504**	**0.633**	16.88	**0.413**	**0.041**
AnchorSplat w/ joint appearance refinement	14.11	0.495	0.651	16.45	0.399	0.044
AnchorSplat w/ unrestricted feature fusion	13.32	0.410	0.674	15.17	0.317	0.049
AnchorSplat w/ render feedback	14.12	0.483	0.640	16.76	0.394	0.042
AnchorSplat w/ auxiliary surfels	13.94	0.453	0.692	**17.03**	0.375	0.048
AnchorSplat, referred scene $\mathcal{G}_q$ only	6.78	0.068	0.908	13.95	0.252	0.108

5.5 Ablation Studies

Core components. With the same DA3 backbone, AnchorSplat surpasses MVGGT+DA3 by 10.1 global mIoU points (Table 5a), attributing the gain to semantic anchoring and geometric transport. The loss caused by substitution is comparable to removing multimodal fusion from MVGGT, which lowers global mIoU from 39.9 to 32.0.

Table 5: Ablations: (a) full-test components (PVSO: per-view no-target suppression); (b) ladder order and post-calibration optimization on fixed subsets.

(a) Components

		Easy		Hard		Overall	
Variant	Backbone	G	V	G	V	G	V
2D-Lift	Pi3	25.4	24.1	6.4	15.0	17.8	20.4
No PVSO, no fusion	Pi3	36.3	43.0	12.9	38.5	26.9	41.1
PVSO only	Pi3	40.7	48.9	19.0	45.4	32.0	47.5
MVGGT	Pi3	50.1	70.6	24.4	**67.3**	39.9	69.3
MVGGT+VGGT-Ω	VGGT-Ω	41.1	60.4	19.7	59.0	32.5	59.9
MVGGT+DA3	DA3	41.3	63.5	15.3	59.7	30.9	62.0
MVGGT+AMB3R	AMB3R	43.0	64.6	17.4	61.1	32.8	63.2
AnchorSplat (full)	DA3	**51.9**	**71.7**	**24.7**	67.1	**41.0**	**69.8**

(b) Alternatives

Variant	Steps	Global	View	A50
(i) Ladder order				
Native order	base	**43.04**	**69.07**	**48.67**
Late local → next global	400	42.47	68.47	46.50
(ii) Post-calibration optimization				
MVGGT+VGGT-Ω (full) (Wang et al., 2026a)	20 epochs	32.54	59.88	32.48
MVGGT+DA3 (full) (Lin et al., 2026)	20 epochs	30.93	62.00	29.24
MVGGT+AMB3R (full) (Wang & Agapito, 2026)	20 epochs	32.79	63.18	31.43
AnchorSplat (calibrated)	0	**41.01**	68.37	**45.00**
+ geometry training	100	40.93	68.09	44.65
+ geometry training	200	40.88	68.05	44.60
+ balanced BCE decoder	100	40.56	68.21	44.65
+ balanced BCE decoder	200	40.31	68.15	44.60
+ boundary refiner	200	**41.01**	**68.39**	**45.00**

Ladder order and optimization. Native block order outperforms a permutation (Table 5b-i). Continued geometry training, loss rebalancing, and boundary refinement do not improve the calibrated model (Table 5b-ii); direct substitutions remain below 33 global mIoU after 20 epochs. This pattern supports the anchored design and is consistent with semantic erosion under continued geometry training.

Fusion stage and decoder depth. Late fusion and 12 multimodal stages perform best (Table 6), since semantic features establish the referent before geometry enters. Every direct substitution scores below all MVGGT variants except the collapsed 16-stage decoder, so replacing the backbone is more

harmful than a suboptimal fusion design. AnchorSplat with late fusion reaches the highest global mIoU of 41.0.

Table 6: Fusion stage and multimodal depth on MVRefer (mIoU, %). MVGGT variants are from Wu et al. (2026).

(a) Fusion stage

	Easy		Hard		Overall	
Fusion	G	V	G	V	G	V
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
MVGGT+VGGT-Ω	41.1	60.4	19.7	59.0	32.5	59.9
MVGGT+DA3	41.3	63.5	15.3	59.7	30.9	62.0
MVGGT+AMB3R	43.0	64.6	17.4	61.1	32.8	63.2
AnchorSplat (late)	**51.9**	**71.7**	**24.7**	67.1	**41.0**	**69.8**

(b) Multimodal stages

	Easy		Hard		Overall	
Stages	G	V	G	V	G	V
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
MVGGT+VGGT-Ω	41.1	60.4	19.7	59.0	32.5	59.9
MVGGT+DA3	41.3	63.5	15.3	59.7	30.9	62.0
MVGGT+AMB3R	43.0	64.6	17.4	61.1	32.8	63.2
AnchorSplat (12)	**51.9**	**71.7**	**24.7**	67.1	**41.0**	**69.8**

5.6 Qualitative Results

Figures 8 and 9 visualize lifted 3D predictions on MVRefer scenes, with point IoU reported for direct substitutions and AnchorSplat.

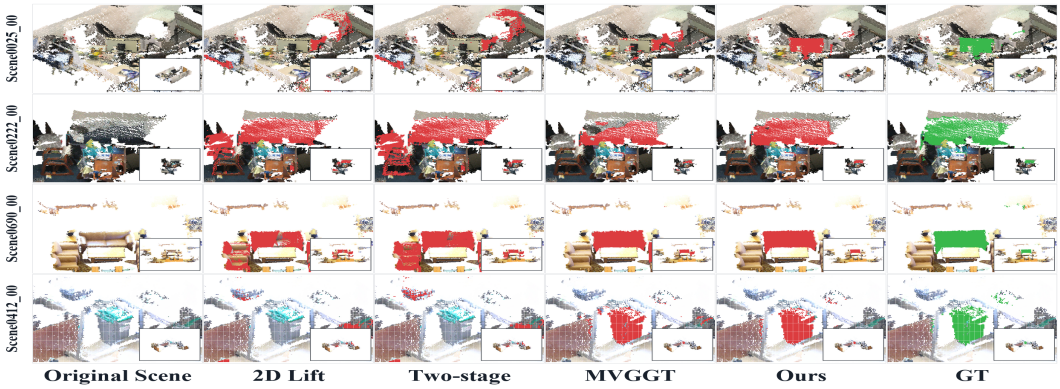


Figure 8: Lifted 3D referring predictions in four MVRefer scenes. Red marks predictions and green marks targets. All methods share eight input views and the visualization geometry.

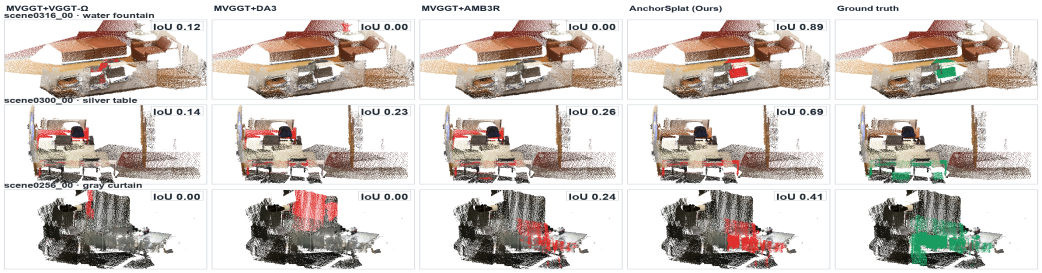


Figure 9: 3D referring predictions of direct substitutions and AnchorSplat. Red marks predictions, green marks targets, and insets report point IoU on the shared canvas.

6 Conclusion

AnchorSplat grounds expressions in a renderable Gaussian scene from sparse uncalibrated views. The semantic anchor counters appearance erosion, geometric correspondences transport referring states, and owner-excluded rendering prevents source-view label copying. On MVRefer, global mIoU rises from 39.9 to 41.0 and Acc@0.50 from 41.5 to 44.6 over MVGGT. One forward pass supports view masks, 3D selection, and novel-view verification.

AI USE STATEMENT

Generative AI assisted with copyediting, LaTeX formatting, and organization of supplementary material. The authors take responsibility for the final manuscript.

ETHICS STATEMENT

This work uses existing indoor scene datasets and collects no new human-subject data. We follow the terms of use of all datasets.

REPRODUCIBILITY STATEMENT

The supplementary material describes the model, training, and evaluation protocols. We will release code and evaluation materials to support reproduction.

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 (ECCV)*, 2020.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 (ECCV)*, 2020.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D Gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision (ECCV)*, 2024.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3772–3782, 2025.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3D. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (ECCV)*, 2024.

Shuting He, Guangquan Jie, Changshuo Wang, Yun Zhou, Shuming Hu, Guanbin Li, and Henghui Ding. ReferSplat: Referring segmentation in 3D Gaussian splatting. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22456–22467, 2025.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

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 (3DV)*, 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 (AAAI)*, 2021.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: Feed-forward 3D Gaussian splatting from unconstrained views. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 44(6), 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, pp. 71–91, 2024.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Xuexun Liu, Xiaoxu Xu, Jinlong Li, Qiudan Zhang, Xu Wang, Nicu Sebe, and Lin Ma. LESS: Label-efficient and single-stage referring 3D segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Alexander Veicht, Sunghwan Hong, Dániel Baráth, and Marc Pollefeys. ZipSplat: Fewer Gaussians, better splats. *arXiv preprint arXiv:2606.05102*, 2026.

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3D reconstruction with backend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. *arXiv preprint arXiv:2605.15195*, 2026a.

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 (CVPR)*, pp. 20697–20709, 2024a.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. EmbodiedScan: A holistic multi-modal 3D perception suite towards embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

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 (ICLR)*, 2026b.

Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. RG-SAN: Rule-guided spatial awareness network for end-to-end 3D referring expression segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 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 (AAAI)*, 2024b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3D Gaussian splats from sparse unposed images. In *International Conference on Learning Representations (ICLR)*, 2025a.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3D scenes. In *European Conference on Computer Vision (ECCV)*, 2024.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025b.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, 2003.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D Gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.