GALE: LEARNING WHERE TO DISTILL FOR SPARSE-VIEW 3D GAUSSIAN SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse-view 3D Gaussian segmentation aims to recover scene geometry and consistent object identities from only a few images. Existing methods use geometry and segmentation foundation models for this task, but do not fully resolve which Gaussians each 2D mask should supervise. This ambiguity arises because Gaussians on different surfaces can project into the same mask. In this paper, we present GALE (Geometry-Aware Lifting of Evidence), a lightweight framework that learns where to distill through geometry-conditioned sparse attention. Built on a VGGT-initialized reconstruction, GALE combines predicted depth and confidence with Gaussian visibility to automatically select and weight the 3D support of each SAM2 mask. These uncertainty-aware assignments focus supervision on compatible Gaussians, guiding knowledge transfer in both scene-specific optimization and cross-scene prediction. Experiments show that GALE outperforms prior distillation and grouping methods on views with held-out annotations. When generalizing to unseen scenes, it surpasses other methods on most metrics using only 1.51M task-specific trainable parameters, approximately 1/40 of the state-of-the-art method.

1 INTRODUCTION

Reconstructing a segmentable 3D scene from a few images requires recovering surfaces and preserving object identities as the camera moves. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) offers a common representation for both: the same primitives can carry appearance, semantic labels, and instance features. Geometry and segmentation foundation models make this goal increasingly accessible. VGGT (Wang et al., 2025) estimates multi-view geometry, while SAM2 (Ravi et al., 2025) identifies image regions. Transferring these complementary priors into a Gaussian scene requires deciding which primitives each region should supervise.

That decision remains ambiguous even when the 2D mask is correct. Gaussians on the visible object, behind it, or near an uncertain boundary can all project inside the mask. Supervising them as one region can therefore pull unrelated features together and mix objects that the teacher separates. Accurate image-space supervision needs a reliable assignment in 3D before it can produce a coherent segmentation. This makes *where to distill* a central question in transferring foundation-model knowledge.

Existing approaches distill dense teacher features into Gaussian fields (Zhou et al., 2024), learn mask-guided affinity or identity features (Cen et al., 2025; Ye et al., 2024), or predict panoptic Gaussians from sparse views (Zhu et al., 2026). Geometry already helps transfer masks and weight feature-to-prototype alignment (Cen et al., 2023; Liao et al., 2026). These methods establish how 2D knowledge can be represented in a renderable field. We focus on the allocation of that knowledge: a mask's support should account for the reliability of predicted geometry, the current scene's visibility, and the region structure learned in its features. Their agreement can guide supervision even when image overlap alone is ambiguous.

We present GALE, *Geometry-Aware Lifting of Evidence*, which learns this assignment through geometry-conditioned sparse attention. Our key insight is that the geometry used to initialize reconstruction can also assess where segmentation evidence is trustworthy as the Gaussian scene evolves. VGGT depth and confidence provide evidence of surface agreement; current Gaussian visibility and learned grouping features refine the assignment. A compact module combines these

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: **GALE constructs a segmentable Gaussian scene.** Representative input views are shown with an RGB rendering and a semantic overlay. The reconstructed geometry is visualized using Gaussian centers, colored by appearance and semantic class.

signals to select support for each SAM2 mask, allowing low-scoring candidates to receive zero weight while retaining graded influence among selected candidates. In scene-specific learning, these weights construct the region prototype used for distillation. Thus the assignment shapes what the features learn, while the region objective trains the assignment itself.

The same principle supports scene-specific optimization and cross-scene prediction, with objectives and decoding adapted to each setting. Within a scene, shared Gaussian support lets observations from different views reinforce one another. Dense label projection complements selective mask votes, preserving semantic coverage beyond their active support. Across scenes, a compact head predicts Gaussian attributes from frozen VGGT features and learns with geometry-conditioned attribution and rendered supervision. It then constructs a segmentable scene from context RGB without per-scene fitting. This extends the role of geometry from recovering a scene to guiding the knowledge transferred into it.

Experiments evaluate whether this transfer remains useful beyond annotated views and training scenes. GALE achieves the highest semantic mIoU among the evaluated distillation and grouping methods on Replica and 3D-OVS; selected views show more continuous regions and less label fragmentation. Trained on ScanNet, it also leads the evaluated feed-forward baselines in semantic mIoU on unseen Replica scenes at two, six, and ten input views. It uses 1.51M task-specific trainable parameters versus EPS3D's 65.60M, roughly one-fortieth as many, excluding frozen backbones and teachers. Controlled ablations further support sparse assignment and dense coverage as complementary parts of the transfer. These results show that learning where supervision acts can connect geometric and segmentation priors effectively within a compact Gaussian framework.

2 Related Work

Foundation-model supervision of Gaussian fields. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) provides explicit primitives that can carry semantic or instance features and render them from new cameras. Feature 3DGS (Zhou et al., 2024) distills dense 2D features through rendered feature maps, while LangSplat (Qin et al., 2024) learns language-aligned Gaussian features for open-vocabulary queries. Region masks from SAM (Kirillov et al., 2023) support more structured supervision. SAGA (Cen et al., 2025) learns scale-gated affinity features for promptable, multi-granularity segmentation; Gaussian Grouping (Ye et al., 2024) learns a compact identity field from SAM masks with 3D spatial consistency. InstanceGaussian (Li et al., 2025) jointly models appearance and semantics before bottom-up instance aggregation. These methods establish Gaussian fields as reusable carriers of 2D knowledge. GALE focuses on how region supervision is distributed over the primitives that contribute to an image.

Geometry-guided mask transfer. Geometric structure helps resolve correspondence across views. SA3D (Cen et al., 2023) transfers masks into a radiance field through density-guided inverse rendering and cross-view self-prompting. SAGD (Hu et al., 2026) identifies and decomposes boundary Gaussians for training-free segmentation. UniSem (Liao et al., 2026) averages region features into prototypes and weights Gaussian-to-prototype alignment by geometric density. GALE learns the support assignment itself: predicted-depth agreement, confidence, current visibility, and Gaussian feature affinity determine sparse weights that also participate in region distillation. Jointly learning

these weights and the Gaussian embeddings allows the allocation of supervision to evolve with the feature field.

Generalizable Gaussian reconstruction. AnySplat (Jiang et al., 2025) predicts Gaussian scenes and camera parameters from uncalibrated images in a single forward pass. For semantic reconstruction, LSM (Fan et al., 2024) maps two unposed images to semantic anisotropic Gaussians; Uni3R (Sun et al., 2026) uses a cross-view transformer to integrate unposed multi-view images. EPS3D (Zhu et al., 2026) predicts semantic and instance features with mutual enhancement for panoptic reconstruction. SpatialSplat (Sheng et al., 2025) separates coarse semantics from fine inter-instance relations to reduce redundancy. UniSem combines error-aware Gaussian dropout with a curriculum of 2D and 3D semantic supervision. These approaches establish generalizable reconstruction and semantic prediction from unposed images. GALE studies geometry-conditioned supervision assignment within this transfer, applying the same principle to scene-specific learning and a compact predictor for unseen scenes.

3 METHOD

GALE determines which Gaussians a 2D mask should supervise, then uses this assignment to learn region features and aggregate segmentation evidence (Figure 2). We develop these three steps for scene-specific learning before describing their cross-scene adaptation in Section 3.3.

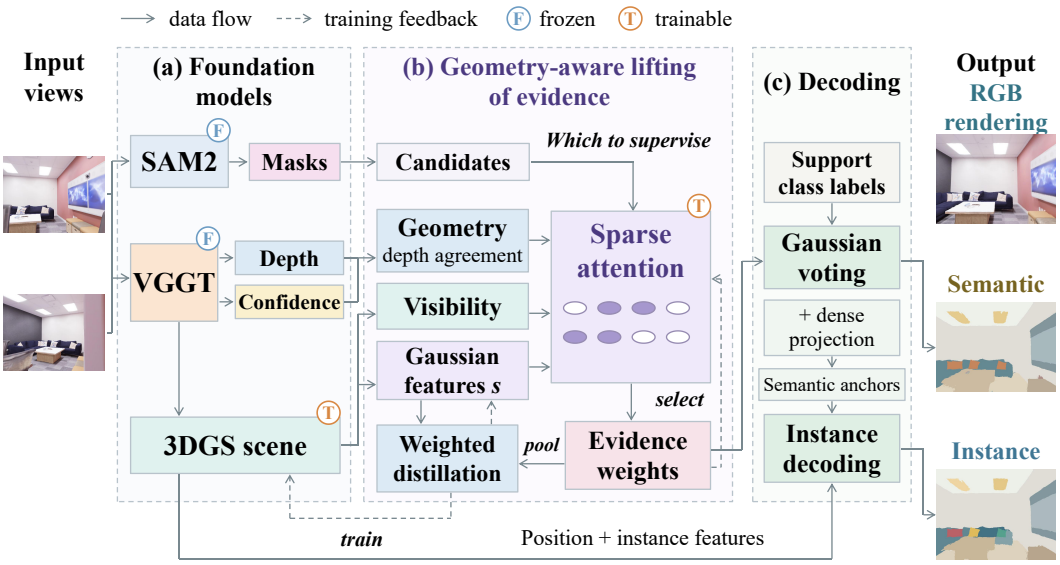


Figure 2: **GALE learns where to distill.** SAM2 masks and VGGT geometry guide sparse selection of learned Gaussian features (content). Selected features form region prototypes for distillation. Votes and dense projection assign classes; positions and instance features separate objects. Hollow markers denote zero weight; solid/dashed arrows show data/training flow. Decoding is scene-specific; attribution extends across scenes. Outputs share a target camera distinct from the illustrated inputs. Its labels are withheld; its RGB may initialize geometry.

3.1 PROBLEM SETUP

We seek a Gaussian scene that renders appearance and segmentation. During fitting or training, $\mathcal{R}$ denotes annotated support views and $\mathcal{T}$ target views reserved for label evaluation. A mask m in view $v \in \mathcal{R}$ supplies evidence; nonnegative weights determine its Gaussian support. Scene-specific fitting learns this assignment within one scene. Cross-scene training instead learns a predictor that constructs scenes from context RGB without annotations or per-scene fitting at inference.

3.2 GEOMETRY-AWARE LIFTING OF EVIDENCE

A shared Gaussian reference. Gaussians provide a common 3D reference for observations across views. Frozen VGGT (Wang et al., 2025) supplies cameras, depth, confidence, and world points; the points initialize or anchor a scene of N Gaussians:

$$\mathcal{G} = \{(\mathbf{x}_i, \Sigma_i, \alpha_i, \mathbf{c}_i, \mathbf{s}_i, \mathbf{z}_i)\}_{i=1}^{N}, \tag{1}$$

where $\mathbf{x}_i$, Σ_i, α_i, and $\mathbf{c}_i$ are Gaussian i's position, covariance, opacity, and appearance. Each Gaussian also carries a learnable grouping embedding $\mathbf{s}_i$, called its *content feature*, and a separate instance embedding $\mathbf{z}_i$. The mask-level compactness and separation objectives in Step 2 train $\mathbf{s}_i$ to represent region membership. Scene-specific fitting optimizes these attributes after VGGT initialization; the amortized model predicts them from VGGT features and world-point anchors.

Step 1: Geometry-aware attribution. A Gaussian inside a mask's projection may lie on another surface. We therefore use projection to define a candidate set $\mathcal{S}_m$: Gaussian indices inside SAM2 mask m that pass image-bound, opacity, and raster-depth checks. Among these candidates, depth agreement indicates surface compatibility, VGGT confidence indicates the reliability of that evidence, and current visibility indicates whether the Gaussian can explain the observation. Using these as graded cues lets the assignment account for uncertain geometry. The descriptor $\mathbf{g}_{im}$ combines them with mask metadata (Appendix A.1).

Geometry can remain ambiguous near boundaries. Learned grouping features provide a complementary test: agreement with the region's current feature representation. We average the candidate embeddings $\mathbf{s}_i$ and project the mean with f_q to obtain a region query $\mathbf{q}_m$; f_k maps each candidate's embedding to a key $\mathbf{k}_i$. Their scaled dot product (Vaswani et al., 2017), termed *content affinity*, measures this agreement. Both inputs come from the learned Gaussian field; SAM2 supplies the masks that define the groups. A small network h adds a geometry score:

$$\mathbf{q}_m = f_q\left(|\mathcal{S}_m|^{-1} \sum_{i \in \mathcal{S}_m} \mathbf{s}_i\right), \qquad \mathbf{k}_i = f_k(\mathbf{s}_i), \tag{2}$$

$$\ell_{im} = \frac{\beta_c \mathbf{q}_m^\top \mathbf{k}_i / \sqrt{d_a} + \beta_g h(\mathbf{g}_{im})}{\tau}, \qquad a_{im} = \mathrm{entmax}_{1.5}(\ell_m)_i, \tag{3}$$

$$w_{im} = |\mathcal{A}_m| a_{im} q_{iv}^{\mathrm{eff}} \alpha_i e^{-4\phi(\Delta d_{iv})}, \qquad \mathcal{A}_m = \{i \in \mathcal{S}_m : a_{im} > 0\}. \tag{4}$$

Here ℓ_m collects logits over $\mathcal{S}_m$, a_{im} is normalized attention, $\mathcal{A}_m$ is its positive support, and w_{im} is the evidence weight. The learned scales β_c, β_g balance feature affinity and geometry; d_a is the query/key dimension and τ the temperature. Entmax-1.5 normalization (Peters et al., 2019) permits exact zero weights, so incompatible candidates need not receive supervision. The factor $|\mathcal{A}_m|$ gives active attention unit mean before reliability weighting, preserving the scale of accumulated votes. Effective VGGT confidence q_{iv}^{eff} and opacity α_i then modulate the selected evidence. The raster-depth gap Δd_{iv} measures distance behind the frontmost projected candidate, and ϕ applies a scaled, clipped logarithmic transform (Appendix A.1). The resulting weight combines learned compatibility with the reliability of the observation and its visibility in the current scene.

Step 2: Weighted region distillation. Attribution and feature learning are coupled through a weighted region prototype. The loss pulls selected Gaussian embeddings toward their prototype, while gradients through its weights train which Gaussians influence it.

In scene-specific learning, an update retains M masks with valid support. For each mask, $\mathcal{B}_m \subseteq \mathcal{A}_m$ is sampled active support, $\bar{w}_{im} = w_{im} / \sum_{j \in \mathcal{B}_m} w_{jm}$ is its normalized weight, and $\hat{\mathbf{s}}_i$ is the unit-length normalization of $\mathbf{s}_i$. The mask prototype $\mathbf{p}_m$ and weighted cosine loss are

$$\mathbf{p}_m = \mathrm{normalize}\left(\sum_{i \in \mathcal{B}_m} \bar{w}_{im} \hat{\mathbf{s}}_i\right), \tag{5}$$

$$\mathcal{L}_{\mathrm{pull}} = \frac{1}{M} \sum_m \sum_{i \in \mathcal{B}_m} \bar{w}_{im} (1 - \hat{\mathbf{s}}_i^\top \mathbf{p}_m). \tag{6}$$

Here $\mathrm{normalize}$ denotes ℓ_2 normalization. The common factor $|\mathcal{A}_m|$ cancels in $\bar{w}_{im}$, leaving relative evidence weights to determine the prototype. Compactness alone admits collapsed features;

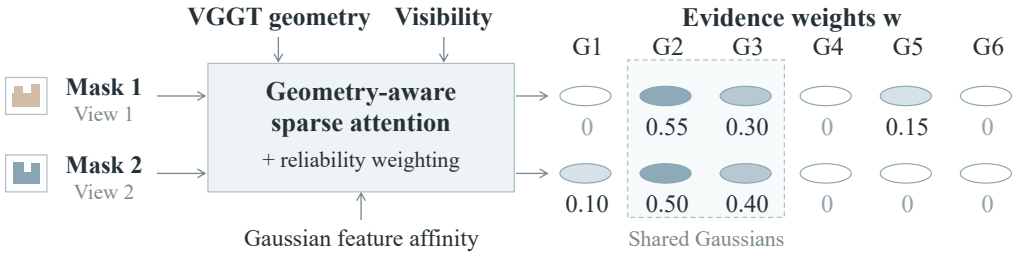


Figure 3: **Geometry-conditioned sparse support.** Masks from different views select support on the same Gaussian set. Geometry, visibility, and Gaussian feature affinity guide sparse attention; reliability weighting produces evidence w_{im}. Hollow markers denote zero weight; boxed columns highlight shared support. Values are illustrative.

prototype separation penalizes agreement between distinct masks, while an entropy target regulates attention concentration. Gradients through positive weights jointly train the Gaussian embeddings and attribution networks f_q, f_k, h. Geometry cues are recomputed from the current scene but detached during correspondence learning, so they guide the grouping objective without being changed to reduce that objective.

Step 3: Evidence aggregation and decoding. Sparse support improves selectivity but need not cover the whole object. In the scene-specific realization, persistent voting combines complementary observations across views, while dense support-label projection supplies semantic evidence beyond active mask support. For task t, an accepted mask-label pair contributes its Step 1 weights to the Gaussian-by-label table $V^{t,\text{mask}}$:

$$V^{t,\text{mask}}_{i,y_t(m)} \mathrel{+}= w_{im}, \qquad t \in \{\text{sem}, \text{inst}\}. \tag{7}$$

Here $V^{t,\text{mask}}$ is a Gaussian-by-label vote table, $y_t(m)$ is the semantic class or cross-view instance ID assigned to mask m, and the increment is applied only to selected support $i \in \mathcal{A}_m$. The semantic label $y_{\text{sem}}(m) = c_m$ is the majority annotated class in the support-view mask; SAM2 supplies no class label. Masks associated across views through shared Gaussian support receive an instance identity $y_{\text{inst}}(m) = k(m)$. The two tables follow the same additive rule. Dense projection adds class evidence on raster-visible Gaussians in V^{dense}, complementing sparse coverage. The combined table gives Gaussian i its semantic prediction $\hat{y}_i$, with c indexing classes:

$$V^{\text{sem}} = V^{\text{sem,mask}} + V^{\text{dense}}, \qquad \hat{y}_i = \arg\max_c V^{\text{sem}}_{ic}. \tag{8}$$

Low-evidence predictions are rejected before label export (Appendix A.1). Dense projection introduces no distillation loss. Normalized semantic votes also anchor instance decoding: high-confidence classes constrain spatial components, and the learned instance field separates objects within them. Accumulated instance votes retain cross-view mask associations separately. Both predictions thus use the same Gaussian scene, with distinct rules for class assignment and object separation (Figure 4).

3.3 Two Operating Regimes

Scene-specific optimization. Scene-specific fitting optimizes VGGT-initialized Gaussian parameters, grouping embeddings $\mathbf{s}$, and instance field $\mathbf{z}$. RGB reconstruction, pixel InfoNCE (van den Oord et al., 2018), and cross-view instance consistency accompany attributed prototype learning (Appendix A.1). The resulting fields render predictions at target cameras without their annotations.

Cross-scene amortization. To transfer without per-scene fitting, a shared head predicts Gaussian appearance, shape, semantic and instance features, and semantic logits from frozen VGGT tokens and world-point anchors. Training retains the sparse attribution operator of Eq. (4). The predicted scene also receives rendered-mask supervision, dense semantic cross-entropy, instance-centroid InfoNCE, and instance-to-semantic consistency. Auxiliary CLIP distillation (Radford et al., 2021) supplies language-aligned features; its teacher and projection are discarded at inference. Prediction then

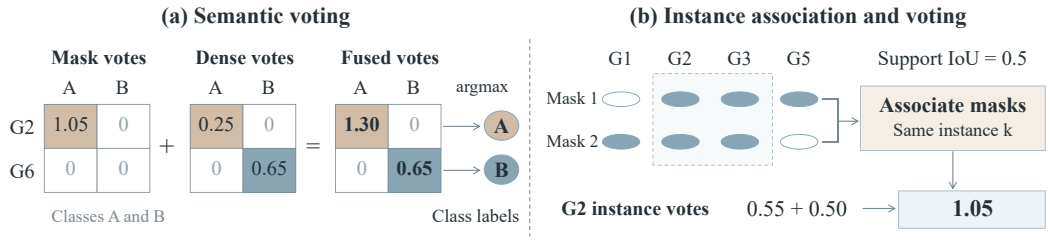


Figure 4: **Scene-specific evidence aggregation.** (a) Mask and dense votes combine before class selection. (b) Shared support associates masks before their weights accumulate as instance evidence. The masks and illustrative weights match Figure 3.

uses context images without optimization or SAM2 masks. The rendered-mask objective supervises each region's pooled semantic prediction. Here, M counts retained training masks, and $\mathcal{P}_m$ indexes pixels in mask m with valid labels and sufficient opacity. Rendered semantic logit vectors $\mathbf{o}_p$ receive opacity-weighted supervision:

$$\mathcal{L}_{\text{mask}} = \frac{1}{M} \sum_m \text{CE}\left(\frac{\sum_{p \in \mathcal{P}_m} \text{sg}(A_p)\mathbf{o}_p}{\sum_{p \in \mathcal{P}_m} \text{sg}(A_p)}, c_m \right), \tag{9}$$

where A_p is opacity at pixel p, c_m is the mask's majority annotated class, CE is cross-entropy, and sg stops gradients through the pooling weights. Appendix A.2 specifies the parameterization, rendering objectives, and loss weights.

4 Experiments

We evaluate GALE in two complementary settings. Scene-specific experiments test whether annotated views support coherent segmentation at other cameras; cross-scene experiments test whether a trained predictor generalizes to unseen scenes without further optimization. We then conduct ablation studies to examine the roles of sparse attribution and evidence aggregation within a scene.

4.1 Experimental Protocols

Scene-specific cross-view transfer. We test annotation transfer on eight Replica (Straub et al., 2019) and ten 3D-OVS (Liu et al., 2023) scenes. Each Replica scene uses 87 annotated support views and 13 uniformly spaced evaluation views ($8 \times 13 = 104$ targets). Each 3D-OVS scene uses two support and three disjoint evaluation views (30 targets). Target annotations never enter fitting or class mapping, but VGGT uses all camera RGB images for geometry estimation. The protocol therefore evaluates label transfer with transductive access to RGB.

To compare region-transfer strategies, we evaluate "SAM2+VGGT," which distills dense SAM2 features into VGGT-initialized Gaussians with the same support masks but no geometry-aware attribution (Appendix A.1); SAGA (Cen et al., 2025), which learns scale-gated grouping affinities; and Gaussian Grouping (Ye et al., 2024), which learns an identity field from SAM masks and spatial consistency. All systems use a common scorer.

Cross-scene transfer to unseen scenes. To assess generalization, we train GALE for 100,000 steps on 1,473 ScanNet (Dai et al., 2017) scenes and evaluate eight unseen Replica scenes without fine-tuning. Budgets $N = 2, 6, 10$ use nested context prefixes and seven fixed queries per scene (56 targets). Context RGB determines the Gaussians; query RGB is used only to estimate cameras, and query labels only for scoring.

We compare three released models: EPS3D (Zhu et al., 2026) predicts semantic and instance Gaussian features; LSM (Fan et al., 2024) reconstructs semantic Gaussians from two unposed images; and Uni3R (Sun et al., 2026) fuses multiple views into a semantic Gaussian scene. All use the same images, scorer, six classes (wall, floor, ceiling, chair, table, sofa), and 448×448 queries. These are system-level comparisons: GALE uses CLIP (Radford et al., 2021) and EPS3D uses LSeg (Li et al.,

2022; Zhu et al., 2026) supervision. Appendix A.2 details rendering, camera estimation, and shared clustering.

Metrics. We report class-wise semantic mIoU, view-averaged mIoU, and mean class recall (class accuracy). Pixel accuracy counts correctly labeled valid pixels; foreground accuracy excludes void/background. Instance mIoU measures matched mask overlap, while F-score@0.5 is precision–recall F1 for matches exceeding mask IoU 0.5, not a geometric distance score. Scene-specific matching is class-gated and scene-global; cross-scene matching is class-agnostic over query views. Instance scores therefore reflect different matching problems.

4.2 Scene-Specific Cross-View Transfer

Table 1: **Scene-specific comparison.** Percentages; "–": unavailable.

Dataset	Method	Sem. mIoU	View mIoU	Class Acc.	FG Acc.	Inst. mIoU	F-score
Replica	SAGA	40.97	43.27	63.87	57.83	–	–
	Gaussian Grouping	26.34	31.76	54.72	49.16	1.74	0.36
	SAM2+VGGT	49.24	56.60	77.28	80.28	–	–
	GALE (ours)	**66.81**	**70.84**	**78.92**	**92.04**	**11.79**	**5.77**
3D-OVS	SAGA	73.97	73.97	90.49	–	–	–
	Gaussian Grouping	79.30	79.30	95.75	–	–	–
	SAM2+VGGT	74.23	74.23	91.03	–	–	–
	GALE (ours)	**90.83**	**90.83**	**96.11**	–	–	–

We first examine whether the lifted supervision remains useful beyond the annotated cameras. GALE achieves the highest semantic mIoU on both datasets (Table 1). Its advantage over SAM2+VGGT, which uses the same support masks, supports learning where those masks should act in 3D. The improvement also holds on 3D-OVS with just two annotated views. In the selected views of Figure 5, GALE produces more continuous object regions with fewer fragmented labels. These examples show corresponding gains in region continuity at unannotated cameras.

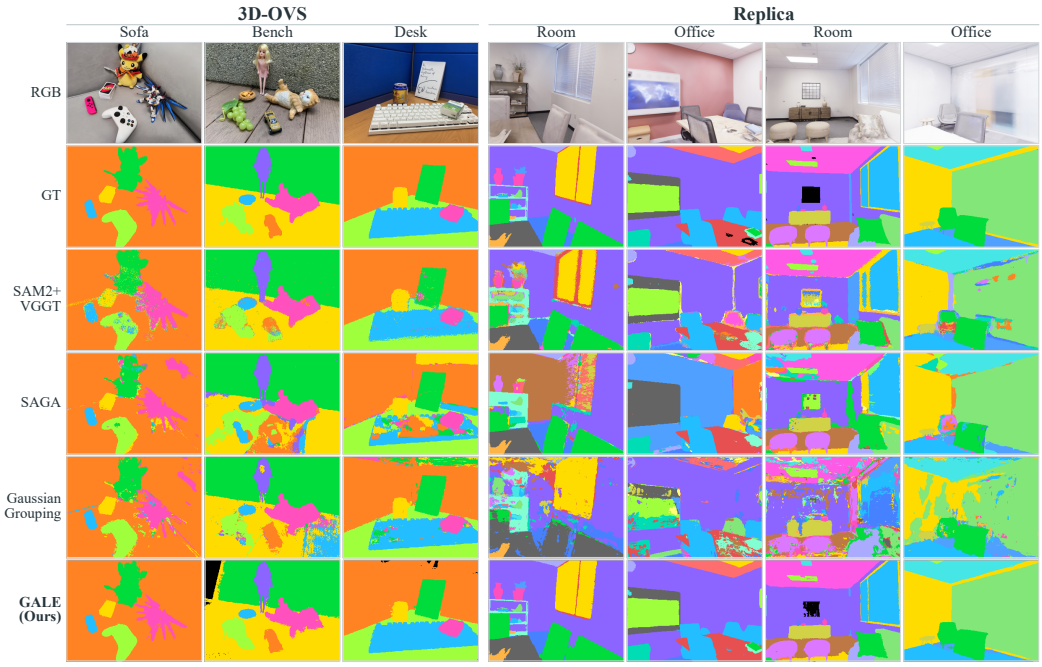


Figure 5: **Cross-view segmentation on 3D-OVS and Replica.** Each column shares a target camera and native resolution. Predictions are unedited and unsmoothed.

4.3 Cross-Scene Transfer to Unseen Scenes

We next ask whether GALE can segment unseen scenes without per-scene fitting. Across all three input budgets, it achieves the highest semantic mIoU and pixel accuracy among the evaluated methods (Table 2). Semantic mIoU rises from 30.78 with two inputs to 51.94 with ten, showing that the same predictor can use additional context at inference. The selected two-view results in Figure 6 further show less wall–floor confusion and a consistent sofa label across the two Office 0 views, illustrating the benefit when observations are sparse.

This semantic accuracy requires only 1.51M task-specific trainable parameters, compared with EPS3D's 65.60M, or approximately one-fortieth as many. Excluding frozen backbones and teachers, this comparison shows that a compact task head can retain strong transfer performance. GALE also leads F-score at each budget, although its instance mIoU is lower than the best baseline; this metric trade-off is examined in Appendix A.3.

Table 2: **Cross-scene evaluation on eight Replica scenes.** Percentages on 56 shared query views per budget. LSM/Uni3R instances use semantic-feature clustering. Bold marks the best per budget.

Views N	Method	Sem. mIoU ↑	Pixel Acc. ↑	Inst. mIoU ↑	F-score@0.5 ↑
2	EPS3D (released)	20.74	35.33	19.76	8.75
	LSM (released)	26.15	73.89	**23.95**	5.79
	Uni3R (released)	29.60	77.68	21.24	5.04
	GALE (ours)	**30.78**	**79.19**	15.26	**21.91**
6	EPS3D (released)	37.28	69.20	**24.80**	7.03
	Uni3R†	36.71	77.35	22.59	4.97
	GALE (ours)	**41.81**	**82.81**	20.92	**15.10**
10	EPS3D (released)	49.40	83.88	**26.49**	6.49
	Uni3R†	34.45	74.49	18.53	5.13
	GALE (ours)	**51.94**	**85.91**	12.22	**19.17**

†Two-view checkpoint evaluated with six or ten inputs. The released LSM implementation accepts two input images; hence it is evaluated only at $N = 2$.

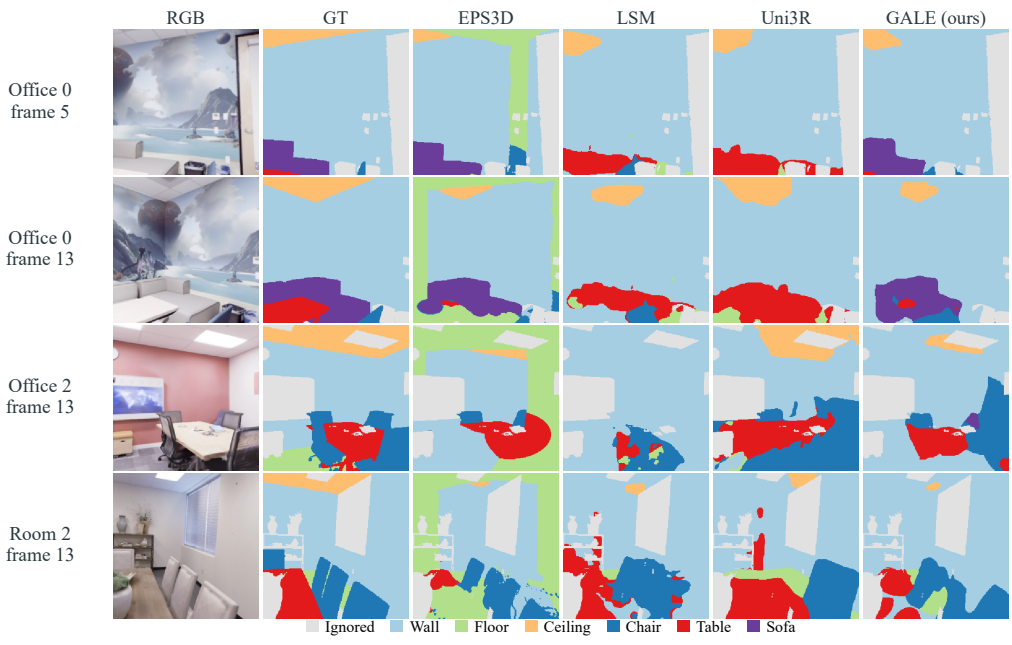


Figure 6: **Two-view cross-scene segmentation on Replica.** Four selected query views, with shared RGB and ground truth. Gray denotes ignored pixels.

4.4 Ablation Study

We conduct three ablations to test the roles of support selection and label coverage: replacing sparse normalization with softmax, removing sparse mask votes, and removing dense projection votes. The variants share ten Replica support frames $(0, 10, \ldots, 90)$, 90 evaluation frames, and seven labeled frames for instance scoring. Table 3 and Figure 7 show their quantitative and visual effects.

Table 3: **Ablation of attribution and evidence aggregation.** Replica 10/90 split, percentages. All rows share the scorer and class-agnostic instance matching.

Variant	Sem. mIoU	FG Acc.	Inst. mIoU
GALE full	**49.24**	**85.70**	**46.55**
w/o sparse voting	48.15	85.49	46.12
Softmax attribution	44.77	85.06	44.21
w/o dense voting	17.25	39.89	32.39

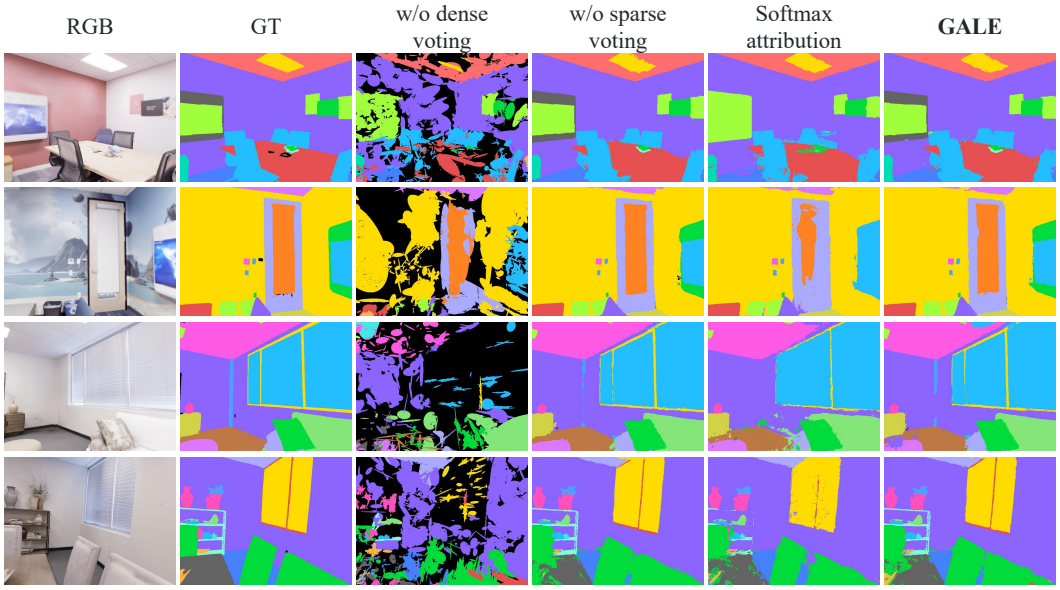


Figure 7: **Selectivity and coverage.** Four selected held-out views from Table 3, with shared row colors. Black denotes ignored ground truth or unassigned predictions.

Softmax reduces semantic mIoU from 49.24 to 44.77. In the selected views, the display surround, door interior, and window divisions are less complete than with the full model. Removing sparse voting has a smaller effect (48.15), suggesting that selective mask votes refine labels supplied by dense projection.

Removing dense voting reduces semantic mIoU to 17.25 and leaves extensive unassigned regions, showing that selective support alone lacks coverage. Dense projection fills these gaps, while sparse attribution helps preserve region structure. The two sources therefore play complementary roles in scene-specific decoding. These studies isolate normalization and aggregation; individual geometric cues and instance association are not separately ablated.

5 Conclusion

GALE learns where 2D masks should supervise a Gaussian scene through geometry-conditioned sparse attribution. This assignment guides region distillation, while complementary evidence aggregation supports scene-specific segmentation. Experiments show improved semantic transfer across views and generalization to unseen scenes with a compact task head, supporting explicit supervision assignment as a way to transfer foundation-model knowledge into segmentable Gaussian scenes.

AI USE STATEMENT

The core research ideas and methodological design were developed by the authors. We used generative AI tools to assist with code implementation based on the authors' design specifications, manuscript editing and language refinement, and retrieval of information on related work and datasets. Additional assistance included preparing explanatory figures, developing supporting scripts, checking implementation descriptions against code, and reviewing and interpreting reported experimental results. The authors take responsibility for the final text, claims, code, and figures.

REFERENCES

Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. Segment anything in 3D with NeRFs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25971–25990, 2023.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3D gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1971–1979, 2025. doi: 10.1609/aaai.v39i2.32193.

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*, pp. 5828–5839, 2017.

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*, volume 37, pp. 40212–40229, 2024.

Xu Hu, Yuxi Wang, Lue Fan, Chuanchen Luo, Junsong Fan, Zhen Lei, Qing Li, Junran Peng, and Zhaoxiang Zhang. SAGD: Boundary-enhanced segment anything in 3D gaussian via gaussian decomposition. *IEEE Transactions on Image Processing*, 35:4934–4946, 2026. doi: 10.1109/TIP.2026.3689408.

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*, 44(6):1–16, 2025. doi: 10.1145/3763326.

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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. InstanceGaussian: Appearance-semantic joint gaussian representation for 3D instance-level perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14078–14088, 2025.

Guibiao Liao, Qian Ren, Kaimin Liao, Hua Wang, Zhi Chen, Luchao Wang, and Yaohua Tang. UniSem: Generalizable semantic 3D reconstruction from sparse unposed images. *arXiv preprint arXiv:2603.17519*, 2026. URL `https://arxiv.org/abs/2603.17519`.

Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3D open-vocabulary segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53433–53456, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, 2019. doi: 10.18653/v1/P19-1146.

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*, pp. 20051–20060, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Yu Sheng, Jiajun Deng, Xinran Zhang, Yu Zhang, Bei Hua, Yanyong Zhang, and Jianmin Ji. SpatialSplat: Efficient semantic 3D from sparse unposed images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26404–26414, 2025.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Xiangyu Sun, Haoyi Jiang, Liu Liu, Seungtae Nam, Gyeongjin Kang, Xinjie Wang, Wei Sui, Zhizhong Su, Wenyu Liu, Xinggang Wang, and Eunbyung Park. Uni3R: Unified 3D reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33280–33290, 2026.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, pp. 5294–5306, 2025.

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*, pp. 20697–20709, 2024.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian Grouping: Segment and edit anything in 3D scenes. In *European Conference on Computer Vision*, pp. 162–179, 2024. doi: 10.1007/978-3-031-73397-0_10.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

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*, pp. 21676–21685, 2024.

Runsong Zhu, Jiaxin Guo, Xiaoyang Guo, Zhengzhe Liu, Ka-Hei Hui, Wei Yin, Kai Chen, Wei Chen, Weiqiang Ren, Yunhui Liu, Pheng-Ann Heng, and Chi-Wing Fu. EPS3D: End-to-end feed-forward 3D panoptic segmentation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

A Implementation Details

A.1 GALE in the Scene-Specific Setting

Geometry descriptor. For candidate i in view v, the geometry descriptor is

$$\mathbf{g}_{im} = [\delta_{iv}, |\delta_{iv}|, q_{iv}, \alpha_i, \phi(\Delta d_{iv}), b_{iv}, \bar{u}_{iv}, \bar{v}_{iv}, \sigma_m, \rho_m], \tag{10}$$

where δ_{iv} compares Gaussian and VGGT depth after separate normalization within view v. Write their depths as d_{iv}^G and d_{iv}^V, with means μ_v^G, μ_v^V and standard deviations s_v^G, s_v^V computed over valid view candidates. With each standard deviation floored at 10^{-6}, the descriptor uses $\delta_{iv} = \frac{1}{4}\,\mathrm{clip}((d_{iv}^G - \mu_v^G)/s_v^G - (d_{iv}^V - \mu_v^V)/s_v^V, -4, 4)$. Raw VGGT confidence Q_{iv} at the projected pixel gives bounded confidence $q_{iv} = Q_{iv}/(Q_{iv}+1)$. The nonnegative raster-depth gap Δd_{iv} is measured from the frontmost projected candidate. Its transform is $\phi(\Delta d) = \frac{1}{4}\,\mathrm{clip}(\log(1 + \Delta d/s_v), 0, 4)$, where s_v is the median positive gap, floored at 10^{-6}, or one if all gaps are zero. The indicator b_{iv} marks valid geometry; fewer than eight valid samples invalidates depth normalization for that view. Normalized coordinates $(\bar{u}_{iv}, \bar{v}_{iv})$ encode image position; σ_m is the SAM2 proposal score and ρ_m its fraction of image area. Invalid VGGT samples receive neutral features and use $q_{iv}^{\mathrm{eff}} = b_{iv}\max(q_{iv}, 10^{-4}) + (1 - b_{iv})$. Thus valid confidence scales the evidence; invalid geometry leaves this multiplier at one so the remaining cues can determine attribution.

Optimization and decoding. The scene-specific objective is

$$\mathcal{L}_{\mathrm{GALE}} = \mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{pull}}\mathcal{L}_{\mathrm{pull}} + \lambda_{\mathrm{push}}\mathcal{L}_{\mathrm{push}} + \lambda_{\mathrm{ent}}\mathcal{L}_{\mathrm{ent}} + \lambda_{\mathrm{inst}}\mathcal{L}_{\mathrm{inst}}. \tag{11}$$

Here $\mathcal{L}_{\mathrm{rgb}}$ is reconstruction loss; $\mathcal{L}_{\mathrm{pull}}$, $\mathcal{L}_{\mathrm{push}}$, and $\mathcal{L}_{\mathrm{ent}}$ are prototype compactness, separation, and attention regularization; $\mathcal{L}_{\mathrm{inst}}$ combines instance supervision terms. The λ coefficients weight these objectives. We optimize each scene for 30,000 iterations, end densification at 15,000, and then run mask and instance updates every ten steps. SAM2 Hiera-L supplies up to 80 proposals per view with at least 35 active Gaussians. Attribution uses temperature 3.0, opacity threshold 0.035, and raster-depth margin 0.015. At most 16 masks and 512 Gaussian embeddings per mask enter each update. Each grouping embedding $\mathbf{s}_i$ has 16 dimensions and is initialized from a standard normal distribution. The projections f_q, f_k each use a bias-free linear layer followed by layer normalization to produce 32-dimensional queries and keys. The attribution module uses AdamW (Loshchilov & Hutter, 2019) (learning rate 10^{-3}, weight decay 10^{-4}); its loss ramps to 0.05 over 1,000 steps, with pull/push/entropy ratios $1.0/0.25/0.01$. The independent 32-dimensional instance field uses AdamW (5×10^{-3}, 10^{-4}). Its outer weight is 0.25 on Replica and 0.10 on 3D-OVS; prototype-pull weights are $0.50/0.25$, respectively, and the semantic-negative weight is 0.10; the prototype cosine margin is 0.30 and the normalized entropy target is 0.35. Replica adds cross-view consistency with weight 0.25. Proposals from different views must have the same semantic label, and their shared-support intersection divided by the smaller support size must be at least 0.02. Within each view, duplicate proposal families use mask IoU 0.75. The Gaussian-by-instance vote table associates supports at IoU 0.45. Both mask vote tables sum accepted updates, including repeated visits to a view. Final instance decoding uses semantic votes normalized over classes as anchor confidence, spatial connectivity, and the independent instance field; the associated instance-vote table is retained separately. Replica semantic export rejects maximum vote scores below two and classes with fewer than 150 assigned Gaussians; unassigned labels remain unknown. Dense projection contributes unit weight per valid Gaussian/view pair. The instance decoder uses anchor confidence 0.80, minimum component size 100, spatial scale 2.5, and minimum final instance size 150.

Sparse normalization. For entmax-1.5 (Peters et al., 2019), a mask-specific threshold κ_m satisfies

$$a_{im} = [(\ell_{im}/2 - \kappa_m)_+]^2, \qquad \sum_{i \in \mathcal{S}_m} a_{im} = 1. \tag{12}$$

Here $(\cdot)_+ = \max(\cdot, 0)$ and κ_m is determined by normalization. In the prototype objective, positive sampled support weights are clamped below at 10^{-8} before renormalization, and feature normalization uses $\epsilon = 10^{-6}$.

Prototype separation and attention regularization. For M masks retained in an update, separation averages over distinct pairs:

$$\mathcal{L}_{\text{push}} = \frac{1}{M(M-1)} \sum_{m \neq n} \left[\mathbf{p}_m^\top \mathbf{p}_n - 0.30\right]_+^2 . \tag{13}$$

It is zero for $M < 2$. Pairs are not filtered by semantic class. The prototypes use the weighted sampled supports in Eq. (6). Attention regularization instead uses the full candidate distribution:

$$H_m = -\frac{\sum_{i \in \mathcal{S}_m} a_{im} \log \max(a_{im}, 10^{-8})}{\log |\mathcal{S}_m|}, \qquad \mathcal{L}_{\text{ent}} = \frac{1}{M} \sum_m (H_m - 0.35)^2. \tag{14}$$

H_m is entropy before active-support subsampling, normalized by the full candidate count. It is set to zero for at most one candidate.

Dense-distillation comparison. The "SAM2+VGGT" comparison fits the same Gaussian scene without geometry-aware mask-to-Gaussian attribution. An MLP maps its low-dimensional semantic field to a 352-dimensional multi-scale SAM2 teacher, supervised by cosine loss. Projected VGGT depth consistency remains in reconstruction. This is a system-level comparison, not an isolated removal of one geometry cue.

Evaluation metrics. For Table 1, we compute IoU for each class in every view where that class is present, average each class over its valid views, and then average classes within a scene; the table macro-averages the resulting scene scores. View mIoU first averages classes within each view. Class accuracy is mean class recall, and foreground accuracy is pixel accuracy after masking void or invalid background. Replica instance scores use class-gated, scene-global Hungarian matching, exclude stuff classes, apply an IoU threshold of 0.5, and macro-average over the eight scenes. On Replica, SAM2+VGGT and SAGA do not export compatible scene-global instance IDs. 3D-OVS provides independent per-class binary masks rather than a mutually exclusive label map, so foreground accuracy and instance metrics are undefined.

A.2 GALE FOR CROSS-SCENE TRANSFER

Model and training. The cross-scene results use the step-100,000 GALE checkpoint trained on 1,473 ScanNet scenes with frozen VGGT and CLIP encoders. AdamW uses learning rate 2×10^{-4}, weight decay 10^{-2}, a cosine schedule, batch size one, and seed 42. The classifier predicts eight ScanNet classes; Replica uses a closed-set argmax over the six shared logits and excludes other ground-truth classes. A biased 1×1 convolution maps 32D semantic features to 512D for CLIP training only. GALE optimizes 1,512,062 task-specific parameters: 1,495,166 in its shared head and 16,896 in this projection. The source-derived accounting of the released EPS3D architecture contains 65,598,657 parameters across its corresponding semantic and instance DPT heads, Sem2Ins fusion, and decoder projection, a 43.4× ratio. At inference, GALE discards the projection. Both counts exclude frozen backbones and teachers.

Gaussian centers combine pooled VGGT world points with a learned residual:

$$\mathbf{x}_i = \text{sg}(\bar{\mathbf{x}}_i^{\text{VGGT}}) + 0.1\, \Delta\mathbf{x}_i. \tag{15}$$

Here $\bar{\mathbf{x}}_i^{\text{VGGT}}$ is a pooled world-point anchor from the current context images and $\Delta\mathbf{x}_i$ the predicted residual; sg stops gradients. The training objective is

$$\mathcal{L} = 0.1\mathcal{L}_{\text{rgb}} + 0.1\mathcal{L}_{\text{sem}} + 0.001\mathcal{L}_{\text{inst}} + 10^{-4}\mathcal{L}_{\text{Ins2Sem}} + \mathcal{L}_{\text{geo}} + 0.1\mathcal{L}_{\text{CLIP}}, \tag{16}$$

where $\mathcal{L}_{\text{rgb}}$ is RGB reconstruction loss, $\mathcal{L}_{\text{sem}}$ averages dense pixel cross-entropy and Eq. (9), $\mathcal{L}_{\text{inst}}$ is instance-centroid InfoNCE, and $\mathcal{L}_{\text{Ins2Sem}}$ enforces instance-to-semantic consistency. $\mathcal{L}_{\text{geo}}$ is smooth L_1 between predicted centers and pooled VGGT points; $\mathcal{L}_{\text{CLIP}}$ is defined below. The CLIP coefficient is fixed, without warm-up or teacher-class agreement gating. Each update samples 2–24 context frames and supervises one rendered context frame.

CLIP distillation. A trainable projection maps the rendered 32-dimensional semantic feature $\mathbf{F}^r$ into the 512-dimensional CLIP space. After resampling to the teacher's 7×7 grid, we minimize

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left[1 - \cos(P_\psi(\mathbf{F}^r)_p, \text{sg}(\mathbf{T}_p))\right], \tag{17}$$

where P_ψ is the projection with learned parameters ψ, $\mathbf{T}_p$ is the frozen CLIP teacher feature at grid location p, and cos is cosine similarity. Ω indexes valid grid locations with rendered opacity above 10^{-3}. Its coefficient is 0.1. Neither the teacher nor the projection is needed at inference.

Mask supervision. We retain masks with at least 16 valid pixels, majority-class purity 0.6, and a majority-to-runner-up margin 0.1; rendered opacity must exceed 10^{-3}. Instance InfoNCE uses SAM2 groups, temperature 0.1, and up to 1,024 pixels. Instance-to-semantic consistency samples 1,000 Gaussian anchors and their 30 nearest neighbors in instance-feature space.

Replica evaluation. We evaluate `office_0–office_4` and `room_0–room_2`, using a fixed frame order per scene. Context candidates have indices $0, 2, \ldots, 18$, and each budget uses a prefix of this order; query indices are $1, 3, \ldots, 13$. All methods use the same images. Query RGB enters only camera estimation, and query labels are loaded only for scoring.

The released LSM and Uni3R checkpoints were trained on ScanNet and ScanNet++ (Yeshwanth et al., 2023). LSM registers RGB-predicted DUSt3R (Wang et al., 2024) point maps against its fixed context scene. Uni3R uses a context-fitted similarity transform and RGB feature/PnP localization, retaining the predicted-camera fallback when geometric checks fail. These RGB-only camera adapters keep every query in the evaluation. No model is fine-tuned on Replica. Uni3R uses its released two-view checkpoint at all budgets. The released LSM implementation jointly processes exactly two input images, with a fixed two-view forward pass and Gaussian head. We therefore evaluate it only at $N = 2$; using six or ten inputs would require an additional multi-view adaptation, such as pairwise reconstruction and fusion.

For LSM and Uni3R, instance IDs are obtained by applying the shared clustering rule to their trained 512-dimensional semantic features. Their feature expansion and bilinear interpolation produce the 448×448 fields. No semantic or alpha foreground mask or target label enters clustering.

All methods render the fixed query views at 448×448. The added baseline renderers use near and far planes of 0.01 and 100, respectively. Semantic logits are interpolated to 256×256 before class selection; scoring excludes a 25-pixel border and unmapped labels. Accuracy is pixel accuracy. Instance features are normalized and clustered jointly across each scene's query views. HDBSCAN (McInnes et al., 2017) uses the same settings for all methods: up to 50,000 sampled features, minimum cluster size 400, and one sample for the core-distance estimate; a single cluster is allowed. Remaining features are assigned to cluster centroids. The sampling seed is 777 plus the scene index. Class-agnostic matching averages pairwise mask IoU over views where the ground-truth instance is visible. Ground-truth stuff and ignored pixels are excluded from IoU. All positive predicted cluster IDs count toward precision, including clusters confined to ignored pixels; no predicted semantic filter is applied. Both instance scores are macro-averaged over all eight scenes, with zero assigned when GT or predictions are empty. The matching definitions appear in Appendix A.3. Source hashes and unrounded measurements are retained with the evaluation records.

A.3 Capacity and Instance Diagnostics

Parameter accounting. Table 4 separates the optimized task-specific modules. GALE's CLIP projection is used only during training, leaving 1,495,166 task-specific parameters at inference. EPS3D's count follows its released perception configuration. Frozen geometry backbones and feature teachers are excluded from both totals; the ratio describes trainable capacity rather than total model size or measured runtime.

Table 4: **Task-specific trainable parameters.** Source-derived counts for the evaluated GALE configuration and released EPS3D configuration.

Method	Module	Parameters
GALE	Shared Gaussian head	1,495,166
	Training-only CLIP projection	16,896
	Total	1,512,062
EPS3D	Semantic DPT head	32,788,256
	Instance DPT head	32,788,256
	Sem2Ins fusion	5,249
	Decoder projection	16,896
	Total	65,598,657

Why the instance metrics differ. With G ground-truth objects and P predicted clusters, Hungarian matching maximizes a $G \times \max(G, P)$ IoU matrix, zero-padded for missing predictions. Instance mIoU averages its G assignments, so extra predictions do not enter the denominator. F-score instead measures object-matching F_1. Only matches with IoU strictly above 0.5 count as TP; precision is TP/P and recall is TP/G. Their harmonic mean is $2\mathrm{TP}/(P+G)$, up to the evaluator's 10^{-6} stabilizer. This penalizes extra predictions.

Table 5 summarizes the saved per-scene matching records at ten context views. GALE produces fewer clusters than EPS3D (30 versus 144), but also slightly fewer matches above IoU 0.5 (5 versus 6). Thus a higher F-score can coexist with lower mean matched IoU. The table contains pooled counts for diagnosis; Table 2 averages scores per scene, so its F-scores cannot be reconstructed from the pooled counts alone.

Table 5: **Instance matching diagnostic at ten context views.** Counts summed over the eight Replica scenes from the same evaluations as Table 2. No prediction or matching threshold is changed.

Method	Predicted clusters	Ground-truth instances	Matches (IoU > 0.5)
EPS3D	144	58	6
GALE	30	58	5