3D-GIA: GEOMETRY-AWARE GLOBAL IDENTITY LEARNING FOR OPEN-VOCABULARY 3D GAUSSIAN SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary 3D segmentation localizes text-specified objects from novel views using a queryable 3D representation. Existing 3D Gaussian Splatting methods often distill view-dependent features and supervise Gaussians with independently generated or tracked masks, causing inconsistent view-local identities and semantic leakage across object boundaries. We propose 3D-GIA, which treats hierarchical masks as unordered observations and associates them in a scene-global identity space. Complementary 3D support, bidirectional reprojection, and appearance propose cross-view matches, while cannot-links provide explicit incompatibility evidence. Confidence-weighted voting projects global clusters to Gaussians as soft identity targets that preserve ambiguous mask-to-3D support. These targets supervise identity and language fields; detached identity predictions guide geometry. 3D-GIA preserves object–part relations, enabling consistent whole-object and part-level queries across views. It achieves 88.0% mIoU and 82.1% mBIoU on LERF-Mask, 96.2% mean mIoU on 3D-OVS, and more consistent cross-view identities than DEVA tracking under a controlled common-mask protocol. The learned identities also support text-guided removal, recoloring, and resizing.

1 INTRODUCTION

Open-vocabulary 3D scene understanding localizes and segments arbitrary targets specified by natural language. CLIP (Radford et al., 2021) and SAM-family models (Kirillov et al., 2023; Ravi et al., 2025) have greatly improved open-vocabulary recognition and class-agnostic segmentation in 2D, motivating their transfer to point clouds, NeRFs, and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023). 3DGS is particularly attractive for its explicit representation and efficient rendering.

However, reliable transfer remains difficult. First, language features are view-dependent, so the same physical object can produce different semantic responses across views. Second, independently generated 2D masks lack consistent identities: one object may fragment into several IDs or merge with its neighbors. Treating these uncertain observations as hard supervision propagates association errors into Gaussian identity and semantic features. Third, photometric 3DGS does not enforce segmentation-aligned boundaries; boundary-spanning Gaussians can therefore leak identity and language features between nearby objects.

The central gap is therefore not merely feature lifting, but the absence of a persistent unit of correspondence. View-local masks and semantic responses cannot jointly supervise a 3D representation unless observations of the same physical entity are reconciled and contradictory observations remain separate. A scene-global identity space provides this missing interface, preserves object–part structure and observation uncertainty, and allows identity learning, semantic transfer, and geometry refinement to share one coherent supervision signal.

We therefore center the framework on scene-global identity. Rather than assigning each view-local mask a brittle track ID, we treat masks as unordered observations in a scene-level graph. Shared 3D support, bidirectional reprojection, and appearance provide complementary *positive evidence* that two observations depict the same entity, whereas explicit *cannot-links* prevent incompatible same-view regions and child–parent-residual pairs from sharing a global identity. Constrained clus-

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.

tering establishes persistent global IDs, and confidence-weighted voting assigns each Gaussian a *soft identity posterior* instead of a one-hot label, preserving ambiguous mask support at the 3D representation level. This posterior is the missing interface identified above: it supervises the identity field and organizes the compact language field, while detached identity predictions guide geometric refinement.

Text queries are then resolved over the persistent identities, which also support removal, recoloring, and resizing (Figure 4).

Our main contributions are summarized as follows:

- We propose 3D-GIA, a conflict-constrained association framework combining 3D support, bidirectional reprojection, appearance, and cannot-links to recover persistent identities from unordered masks and project mask-to-3D uncertainty to Gaussian soft posteriors.
- We construct a hierarchy-aware mask bank that preserves object–part relations for conflict construction and consistent whole/part queries.
- We use the shared identity space to stabilize language learning and guide boundary-aware geometry, producing more coherent semantics and thinner, boundary-aligned Gaussians.
- Extensive experiments demonstrate strong open-vocabulary segmentation and reliable cross-view identity association. The learned scene-global identities also enable precise text-guided 3D scene editing.

2 Related Work

Open-Vocabulary 3D Understanding. Point-cloud and multimodal methods lift 2D cues into queryable 3D features or instances (Jatavallabhula et al., 2023; Peng et al., 2023; Takmaz et al., 2023; Nguyen et al., 2024), while LERF and 3D-OVS embed language features in radiance fields (Kerr et al., 2023; Liu et al., 2023). Gaussian methods distill hierarchical or granularity-aware features (Qin et al., 2024; Peng et al., 2026), aggregate them over learned instance regions (Piekenbrinck et al., 2025), or improve language alignment with instance identities and cross-view context (Jang & Kim, 2025; Huang et al., 2025). In contrast, we construct a scene-global identity space from unordered masks and use confidence-weighted cluster support to supervise both identity and language fields.

Multi-View Mask Association and Gaussian Grouping. Gaussian Grouping obtains identity supervision from tracker-associated SAM masks (Ye et al., 2024), while SAGA learns scale-gated affinities from multi-granularity masks without persistent observation identities and OmniSeg3D learns hierarchical cross-scale features (Cen et al., 2025; Ying et al., 2024). Gaga incrementally reconciles inconsistent masks with a 3D-aware memory bank (Lyu et al., 2026), and Proto-SaGa combines semantic-aware association with prototype classification (Oh et al., 2026). Trace3D records per-view instance evidence to correct masks and refine ambiguous Gaussians (Shen et al., 2025), while concurrent PairGS predicts pairwise Gaussian relations and builds a hierarchical cluster tree (Cha et al., 2026). In contrast, 3D-GIA performs conflict-constrained global clustering over coarse-to-fine *mask observations* using shared 3D support, bidirectional reprojection, appearance, and hierarchy-derived conflicts. The resulting soft posteriors and positive/cannot-link pairs supervise identity and language learning, while detached identity predictions guide geometry.

Geometry-Aware Gaussian Segmentation. Surface methods regularize Gaussian geometry using mesh alignment, depth/normal priors, opacity fields, or planar constraints (Guédon & Lepetit, 2024; Huang et al., 2024; Turkulainen et al., 2025; Yu et al., 2024; Chen et al., 2025); semantic methods split or prune ambiguous primitives using boundary gradients or instance evidence (Zhang et al., 2025; Shen et al., 2025). Our term instead continuously adjusts positions, scales, and rotations using local surface affinity and detached identity discrepancy, while flattening limits thick or wide support. Thus geometry is coupled to the identities used for inference without a separate boundary detector or discrete split/prune decisions.

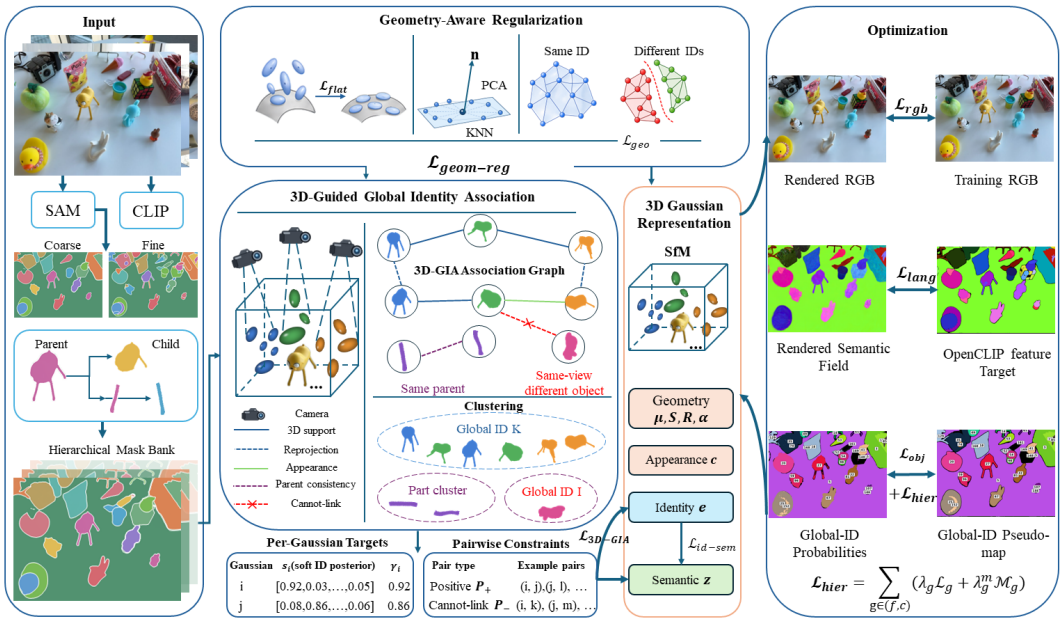


Figure 1: Overview of 3D-GIA. Hierarchical view-local masks are associated into persistent scene-global identities. Their Gaussian-level soft posteriors supervise identity and language learning, while detached identity predictions guide geometry.

3 METHOD

3.1 REPRESENTATION AND OVERVIEW

Gaussian fields. Given posed images, 3DGS represents a scene by anisotropic Gaussians $\mathcal{G} = \{g_i\}_{i=1}^{N}$ with center $\boldsymbol{\mu}_i$, covariance $\boldsymbol{\Sigma}_i$, opacity α_i, and view-dependent color $\mathbf{c}_i$. Any per-Gaussian attribute $\mathbf{x}_i$ is rendered at pixel $\mathbf{p}$ in view v by standard front-to-back alpha compositing:

$$\mathbf{X}_v(\mathbf{p}) = \sum_{i \in \mathcal{I}_v(\mathbf{p})} T_{i,v}(\mathbf{p}) a_{i,v}(\mathbf{p}) \mathbf{x}_i. \tag{1}$$

Here $\mathcal{I}_v(\mathbf{p})$ is the depth-ordered set of contributing Gaussians, while $a_{i,v}$ and $T_{i,v}$ are projected opacity and accumulated transmittance. Each Gaussian additionally carries a 16-D identity embedding $\mathbf{e}_i$ and a 3-D language code $\mathbf{z}_i$. For 2D language supervision, preprocessing extracts a normalized OpenCLIP descriptor for each SAM-derived image segment. An offline autoencoder compresses these descriptors to 3-D, and the segment-index map at the selected granularity broadcasts each code to its pixels. The frozen decoder maps a code back to the OpenCLIP space. The autoencoder is trained with $\mathcal{L}_{AE} = \mathcal{L}_2 + \lambda_{AE}\mathcal{L}_{cos}$, where $\mathcal{L}_2$ and $\mathcal{L}_{cos}$ are squared-error and cosine losses, and $\lambda_{AE} = 0.01$. The same compositing renders RGB $\mathbf{C}_v$, identities $\mathbf{E}_v$, and language codes $\mathbf{Z}_v$.

Pipeline. As shown in Figure 1, 3D-GIA turns independently generated 2D masks into persistent, queryable 3D units in five stages. First, coarse and fine masks are decomposed into atomic observations while retaining their object–part hierarchy. Second, a warm-up 3DGS lifts these observations into 3D, where complementary support, reprojection, and appearance cues propose cross-view matches and cannot-link constraints suppress incompatible merges, consolidating view-local IDs into persistent scene-global identities. Third, the resulting global clusters are projected to Gaussians as confidence-weighted soft identity posteriors, preserving ambiguous mask support instead of forcing brittle hard labels. Fourth, this shared posterior supervises a renderable identity field and organizes the language field through identity-conditioned prototypes; the learned identity field then supplies detached differences for boundary-aware geometry refinement. Finally, open-vocabulary queries retrieve persistent global identities rather than unrelated pixel responses across views.

3.2 Hierarchy-Aware Mask Observations

Coarse masks capture whole objects, whereas fine masks recover parts; directly associating their overlapping pixels would confuse these two roles. We use SAM2 (Ravi et al., 2025) to generate sparse-prompt coarse masks and dense-crop fine masks. When a retained fine mask is contained by a coarse mask, we assign it to the best-containing coarse mask, giving a parent–child edge. Before association, each coarse parent is replaced by its *parent residual*, obtained by removing the union of its children; an empty residual is discarded. Parent residuals and fine masks are therefore separate observations, while their hierarchy is retained.

This distinction is used consistently in association and training, and can also support structured querying. A child and its parent residual are cannot-linked so that their global identities remain distinct. The original hierarchy bank also supplies child-region and parent-residual targets for identity learning. After association, each retained parent–child observation relation is mapped to the corresponding global-ID relation. These relations can be exported as an optional query hierarchy: a parent selection may include its descendant IDs, whereas a child selection remains part-specific (Figure 2).



Figure 2: Optional hierarchy-aware querying expands a parent residual to its children for whole-object queries, while part queries remain child-specific.

3.3 Conflict-Constrained Association

Observation graph. Global association needs both supporting and prohibitive evidence. No single positive cue is reliable: 3D support can merge touching objects, appearance can confuse similar instances, and reprojection can fail under occlusion. We therefore use these cues as complementary evidence for a match, while representing known incompatibilities with explicit cannot-link constraints. We voxelize warm-up Gaussian centers into support primitives q_r using COLMAP visibility (Schönberger & Frahm, 2016). For observation n with mask M_n in view v, $\mathcal{Q}_n$ contains the primitives whose visible projections fall inside M_n; thus, the standard posed-RGB setting uses no external depth maps. Each observation retains its mask, 3D support, confidence, pooled OpenCLIP descriptor, and parent relation. For observations a and b, the continuous edge score e_{ab} ranks matches by 3D-support overlap, bidirectional reprojection overlap, and appearance cosine similarity. The three cues are not required to be simultaneously strong: appearance acts as a semantic gate, while 3D overlap and reprojection provide alternative geometric support. Thus, an ordinary pair enters the initial union pass only when appearance agrees and at least one geometric cue passes its threshold; e_{ab} orders admitted edges using the combined cue score. We set $m_{ab} = 1$ when this positive-evidence gate is satisfied. A same-view exception is allowed only when appearance, nearby 3D support, and shared-parent evidence agree. Cannot-links, introduced next, can veto either type of positive evidence.

We set $c_{ab} = 1$ for a final cannot-link and collect these pairs in $\mathcal{N} = \{(a, b) : c_{ab} = 1\}$. They include incompatible same-view regions, a direct child–parent-residual pair, observations with incompatible verified parents, or mutually visible observations with inconsistent reprojection. Two components are compatible only if no member pair has a cannot-link and they contain multiple observations from one view only when their same-view relation was verified. After edge-level unions, a component pair is considered again using the mean of its strongest k_C cross-component edges and is merged when this mean exceeds τ_C. This procedure is independent of temporal ordering. Exact tests, k_C, and thresholds are given in the supplement.

Gaussian-level soft identity posterior. The graph yields discrete observation clusters, but their transfer to 3D is uncertain because visibility, overlap, and mask noise provide different amounts of evidence to each primitive. Confidence-weighted voting preserves this ambiguity instead of forcing a one-hot target. Let $\mathcal{Q}_n$ be the 3D support of observation n, and let $\beta_n \in [0, 1]$ be its effective vote weight. It combines SAM mask quality, valid projection/visibility support during lifting, and the reliability of its cluster. For primitive q_r, the evidence of global cluster $\mathcal{C}_k$ is

$$u_{r,k} = \sum_{n \in \mathcal{C}_k} \beta_n \mathbf{1}[q_r \in \mathcal{Q}_n]. \tag{2}$$

Here $\mathbf{1}[\cdot]$ is the indicator. For primitives with nonzero total evidence, negligible identity hypotheses are pruned and normalization yields the posterior and its concentration:

$$s_{r,k} = \frac{u_{r,k}}{\sum_\ell u_{r,\ell}}, \qquad \gamma_r = \|\mathbf{s}_r\|_\infty. \tag{3}$$

Here $\mathbf{s}_r$ describes ambiguity among the fixed global clusters at a primitive, and γ_r measures posterior concentration rather than total evidence mass; it is not a posterior over alternative graph partitions. Unsupported primitives receive no soft supervision, and primitives not above the concentration threshold τ_{id} are excluded from the soft loss. Nearest-center transfer gives Gaussian posteriors and concentrations $(\mathbf{s}_i, \gamma_i)$. We sample positive primitive pairs from the same cluster and negative pairs across cannot-linked or incompatible co-visible clusters, then map their endpoints to the nearest Gaussians to obtain $\mathcal{P}^+$ and $\mathcal{P}^-$. Finally, assigning each observation's cluster ID to its pixels, with background ID zero and finer regions taking precedence, produces the dense maps $\mathbf{Y}_v$ used below.

3.4 Global Identity and Language Learning

Offline clusters cover only observed masks. We therefore learn a renderable per-Gaussian identity field from their soft targets and reuse this identity space to make language features agree within each physical entity. **Identity supervision.** Let h_ϕ be a fixed-capacity identity classifier with C outputs, where zero denotes background and the K global clusters occupy foreground labels; we choose $C \geq K + 1$, leaving any remaining outputs unused. Its rendered and per-Gaussian probabilities are $\mathbf{P}_v = \mathrm{softmax}(h_\phi(\mathbf{E}_v))$ and $\mathbf{p}_i = \mathrm{softmax}(h_\phi(\mathbf{e}_i))$, respectively. To match this fixed label space, $\mathbf{s}_i^+ \in \mathbb{R}^C$ denotes $\mathbf{s}_i$ augmented with zero background mass and zero entries for unused foreground outputs. The dense maps $\mathbf{Y}_v$ use area-balanced cross-entropy $\mathcal{L}_{obj}$. For hierarchy level $g \in \{f, c\}$, where f and c denote fine children and coarse-parent residuals, region-balanced $\mathcal{L}_g$ and child–parent margins $\mathcal{M}_g$ give $\mathcal{L}_{hier} = \sum_g (\lambda_g \mathcal{L}_g + \lambda_g^m \mathcal{M}_g)$. The margins require the child identity to dominate on child pixels and the parent identity to dominate on its residual. Association pairs additionally pull matched identities together and separate cannot-linked identities. For embeddings $\mathbf{x}$, margin m, and $[t]_+ = \max(t, 0)$, the pair loss is

$$\ell_{pair}(\mathbf{x}; \mathcal{P}^\pm, m) = \mathbb{E}_{\mathcal{P}^+}[1 - \cos(\mathbf{x}_i, \mathbf{x}_j)] + \mathbb{E}_{\mathcal{P}^-}[\cos(\mathbf{x}_i, \mathbf{x}_j) - m]_+. \tag{4}$$

Together with soft cross-entropy on Gaussians having sufficiently concentrated posteriors, $\mathcal{V} = \{i : \gamma_i > \tau_{id}\}$, the identity objective is

$$\mathcal{L}_{3D\text{-}GIA} = \lambda_{soft}\, \mathrm{CE}_{\mathcal{V}}(\mathbf{s}^+, \mathbf{p}) + \lambda_{pair} \ell_{pair}(\mathbf{e}; \mathcal{P}^\pm, m_{id}). \tag{5}$$

This association-specific objective complements the dense and hierarchy terms; the complete identity supervision is $\lambda_{obj}\mathcal{L}_{obj} + \mathcal{L}_{hier} + \mathcal{L}_{3D\text{-}GIA}$. Exact hierarchy normalization and margins are given in the supplement.

Language distillation and identity transfer. Each rendered language-code map first matches the corresponding frozen autoencoder code with an ℓ_1 loss; the decoder is used to recover OpenCLIP features for text similarity at inference. This view-wise distillation alone does not make the same object semantically consistent across views, so we use global identity as a transfer scaffold. The fixed association posterior provides stable scene-level supervision, while the learned identity field provides renderable evidence; their agreement forms a reliable transfer teacher.

With Norm_p denoting ℓ_p normalization, $\odot$ elementwise multiplication, and sg stop-gradient, their product forms a detached agreement teacher:

$$\mathbf{r}_i = \mathrm{Norm}_1\big(\mathbf{s}_i^+ \odot \mathrm{sg}(\mathbf{p}_i)\big), \quad w_i = \mathrm{sg}\big(\gamma_i \|\mathbf{p}_i\|_\infty (\mathbf{s}_i^+)^\top \mathbf{p}_i\big). \tag{6}$$

Let $\mathcal{V}_{\mathrm{sem}}$ be Gaussians passing the posterior-concentration and identity-prediction-confidence thresholds and having nonzero fixed/learned identity agreement; exact thresholds are given in the supplement. The detached weight w_i combines posterior concentration, prediction confidence, and their agreement. Their normalized codes $\bar{\mathbf{z}}_i = \mathrm{Norm}_2(\mathbf{z}_i)$ are aggregated into identity prototypes. Let $\mathcal{K}^+$ contain identities with sufficient teacher support. For $k \in \mathcal{K}^+$,

$$\boldsymbol{\pi}_k^{\mathrm{sem}} = \mathrm{sg}\left[\mathrm{Norm}_2\left(\sum_{i \in \mathcal{V}_{\mathrm{sem}}} w_i r_{i,k} \bar{\mathbf{z}}_i\right)\right]. \tag{7}$$

Let $\widehat{\mathbf{r}}_i$ be $\mathbf{r}_i$ restricted and renormalized over $\mathcal{K}^+$. Each Gaussian then receives the resulting prototype mixture:

$$\widetilde{\boldsymbol{\pi}}_i^{\mathrm{sem}} = \mathrm{Norm}_2\left(\sum_{k\in\mathcal{K}^+} \widehat{r}_{i,k}\boldsymbol{\pi}_k^{\mathrm{sem}}\right). \tag{8}$$

The compactness term pulls $\bar{\mathbf{z}}_i$ toward $\widetilde{\boldsymbol{\pi}}_i^{\mathrm{sem}}$; a soft classification term matches prototype similarities to $\widehat{\mathbf{r}}_i$; and the positive/cannot-link pair loss is also applied to language codes. Weighted by λ_{cmp}, λ_{ce}, and λ_{sp}, respectively, these terms form $\mathcal{L}_{id\text{-}sem}$. All targets are built only from training views and are stop-gradient, preventing semantic learning from distorting the identity predictor. The exact normalization is given in the supplement. We denote by $\mathcal{L}_{part}$ an object-internal part-preservation term. Within each supervised global-identity region, the SAM-derived segments used for language targets define parts whose mean codes are aligned to their frozen segment features, while different parts are separated. These segments provide within-region language supervision and do not introduce a second cross-view identity space.

3.5 Identity-Guided Geometry Regularization

Feature supervision alone cannot remove leakage when a Gaussian physically spans two objects, so this module changes the representation itself. An identity-free term flattens thick or wide Gaussians; an identity-guided term keeps local geometry continuous within an identity and separates it across identity boundaries. Identity predictions are detached so that they guide geometry rather than adapt to make the geometry loss easier. Let $\rho_i^{min} \le \rho_i^{mid} \le \rho_i^{max}$ be the sorted Gaussian axis scales. On a directed k-NN graph, $\mathbf{n}_i$ fuses the minimum-scale Gaussian normal with a local PCA normal. Normal, point-to-plane, and spatial deviations $\delta_{ij}^n = 1 - |\mathbf{n}_i^\top \mathbf{n}_j|$, $\delta_{ij}^p = |(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^\top \mathbf{n}_i|$, and $\delta_{ij}^x = \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|_2^2$ define the local surface affinity

$$A_{ij} = \mathbf{1}\left[|\mathbf{n}_i^\top \mathbf{n}_j| \ge \cos\theta_g\right] \cdot \exp\left(-\frac{\delta_{ij}^n}{\sigma_n} - \frac{\delta_{ij}^p}{\sigma_d} - \frac{\delta_{ij}^x}{\sigma_x}\right). \tag{9}$$

Here $\theta_g = 40^\circ$ is the normal gate, and the positive $\sigma_n, \sigma_d, \sigma_x$ are affinity bandwidths. For the directed edge set $\mathcal{E}_k$, let $\langle x_{ij}\rangle_w = \sum_{(i,j)\in\mathcal{E}_k} w_{ij}x_{ij} / (\sum_{(i,j)\in\mathcal{E}_k} w_{ij} + \epsilon)$ for $\epsilon > 0$, and $D_{ij} = \|\,\mathrm{sg}(\mathbf{p}_i) - \mathrm{sg}(\mathbf{p}_j)\|_2$. We penalize excessive width and footprint by $r_i^s = [\rho_i^{max} - \tau_s]_+^2$ and $r_i^a = [\rho_i^{mid}\rho_i^{max} - \tau_a]_+^2$, where τ_s, τ_a are fixed shape thresholds. With identity margin m_{geom}, the complete weighted geometry objective is

$$\mathcal{L}_{geom\text{-}reg} = \frac{\lambda_{flat}}{N}\sum_i(\rho_i^{min} + \lambda_s r_i^s + \lambda_a r_i^a) + \lambda_{cont}\langle D_{ij}^2\rangle_A + \lambda_{sep}\langle[m_{\mathrm{geom}} - D_{ij}]_+\rangle_{1-A}. \tag{10}$$

The five λ coefficients weight, in order, flattening, scale, footprint, within-identity continuity, and cross-identity separation. Since D_{ij} is detached, this objective updates only Gaussian positions, scales, and rotations. The first line flattens Gaussians. In the second line, a large identity discrepancy penalizes high surface affinity, whereas a small discrepancy penalizes low affinity; the geometry therefore separates across identity boundaries while remaining continuous within an identity.

3.6 Training and Open-Vocabulary Inference

After an 8K warm-up of geometry, opacity, RGB, and language, we run offline 3D-GIA and optimize to 24K. $\mathcal{L}_{obj}$ begins after association; $\mathcal{L}_{3D\text{-}GIA}$ and $\mathcal{L}_{hier}$ start at 12K, $\mathcal{L}_{id\text{-}sem}$ at 14K, and $\mathcal{L}_{part}$ and $\mathcal{L}_{geom\text{-}reg}$ at 16K; identity-guided geometry is warmed through 20K.

At inference, the frozen decoder maps $\mathbf{Z}_v$ to normalized OpenCLIP features, and $\arg\max_k P_{v,k}$ defines identity regions. Each foreground identity is scored by the mean of its top κ pixel–text similarities. For 3D-OVS, we select view-wise at most eight identities scoring at least τ_{rel} times the best and threshold their summed probabilities at τ_m. For LERF-Mask, first-view Grounded-SAM query masks (Liu et al., 2024; Kirillov et al., 2023) select overlapping identities once, which remain fixed across views. Thus, benchmark-specific selection differs while the learned identities remain persistent across views. Optional structured querying expands a selected parent to its descendants. Exact thresholds are in the supplement.

Table 1: Results on LERF-Mask (%).

Method	ramen		figurines		teatime		Avg.	
	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU
LangSplat (Qin et al., 2024)	50.4	44.7	52.8	50.5	69.5	65.6	57.6	53.6
Gaussian Grouping (Ye et al., 2024)	77.0	68.7	69.7	67.9	71.7	66.1	72.8	67.6
OpenSplat3D (Piekenbrinck et al., 2025)	75.9	68.2	**92.3**	89.4	83.7	78.8	84.0	78.8
ILGS (Jang & Kim, 2025)	84.3	75.5	75.9	73.8	81.2	78.8	80.5	76.0
Proto-SaGa (Oh et al., 2026)	75.6	69.7	88.5	86.1	79.3	75.5	81.1	77.1
Gaga (Lyu et al., 2026)	72.0	63.3	**92.3**	**90.8**	71.2	68.4	78.5	74.2
COB-GS (Zhang et al., 2025)	78.1	69.2	76.3	73.9	77.2	72.8	77.2	72.0
ObjectGS† (Zhu et al., 2025)	73.9	65.7	89.6	86.7	72.8	69.2	78.8	73.9
3D-GIA (**Ours**)	**88.8**	**78.5**	89.4	87.1	**85.9**	**80.7**	**88.0**	**82.1**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. For open-vocabulary segmentation, we optimize one model per LERF-Mask (Ye et al., 2024; Kerr et al., 2023) and 3D-OVS (Liu et al., 2023) scene using posed RGB images only. Following established protocols, LERF-Mask fixes the scene-global identities selected by Grounded-SAM query masks from the first evaluation view, whereas 3D-OVS performs view-wise language retrieval (Qin et al., 2024). Controlled association instead uses Gaga's seven-scene ScanNet subset (Dai et al., 2017; Lyu et al., 2026) with dataset-provided depth and camera calibration: 639 views and 189 recurring instances. Scene details are in the supplement.

Baselines. We compare with LangSplat, Gaussian Grouping, SAGA, OpenSplat3D, ILGS, COB-GS, ObjectGS, Gaga, and Proto-SaGa (Qin et al., 2024; Ye et al., 2024; Cen et al., 2025; Piekenbrinck et al., 2025; Jang & Kim, 2025; Zhang et al., 2025; Zhu et al., 2025; Lyu et al., 2026; Oh et al., 2026). Reported values follow official protocols; † denotes released-code reproductions. Gaussian Grouping constructs cross-view mask IDs through the SAM–DEVA pipeline (Ye et al., 2024). We therefore use DEVA's native RGB temporal consensus and propagation with externally supplied masks (Cheng et al., 2023), termed *DEVA tracking*, to test whether calibrated geometry and global conflicts provide better identity supervision for posed reconstruction.

Metrics. We report mean intersection over union (mIoU) on both benchmarks and mean boundary IoU (mBIoU) on LERF-Mask. On ScanNet, association accuracy (AssA), ground-truth-balanced association quality (AQ), IDF1, and B^3-F1 measure association and clustering. Sw. is the total number of predicted-identity changes between consecutive observations of the same ground-truth instance; Fr. is the mean number of additional predicted identities assigned to each recurring ground-truth instance; and Cov. is the mean fraction of ground-truth foreground pixels assigned a predicted identity. All metrics use only predicted/ground-truth identity assignments.

Implementation Details. We use SAM2.1 Hiera-L (Ravi et al., 2025) and LAION-2B OpenCLIP ViT-B/16 (Cherti et al., 2023). The dense identity and language-reconstruction losses use unit weights. We set $(\lambda_f, \lambda_c, \lambda_f^m, \lambda_c^m) = (.18, .12, .04, .03)$, $(\lambda_{soft}, \lambda_{pair}) = (.10, .05)$, $(\lambda_{cmp}, \lambda_{ce}, \lambda_{sp}) = (.030, .0125, .010)$, and $(\lambda_{flat}, \lambda_{sep}, \lambda_{cont}) = (.06, .15, .20)$. The segment-alignment and within-object separation terms in $\mathcal{L}_{part}$ are weighted $(.04, .01)$, with $\lambda_s = \lambda_a = 1$. The RGB DSSIM mixture is .2, and the offline autoencoder uses $\lambda_{AE} = .01$ with a 3-D bottleneck. Each scene is trained for 24K iterations on one NVIDIA RTX A6000; thresholds and other details are in the supplement.

4.2 QUANTITATIVE RESULTS

LERF-Mask. Table 1 reports the strongest average performance under the official protocol, especially with adjacent distractors and fine-grained parts.

3D-OVS. Table 2 reports the best mean mIoU across varied layouts, supporting identity-level retrieval.

Table 2: mIoU (%) on 3D-OVS.

Method	bed	bench	room	sofa	lawn	Mean
LangSplat	92.5	94.2	94.1	90.0	96.1	93.4
GS Grouping	83.0	91.5	85.9	87.3	90.6	87.7
ILGS	96.4	95.5	93.6	92.2	94.2	94.4
SAGA	**97.4**	95.4	96.8	93.5	96.6	96.0
3D-GIA (**Ours**)	94.9	**95.8**	**97.7**	**95.5**	**97.1**	**96.2**

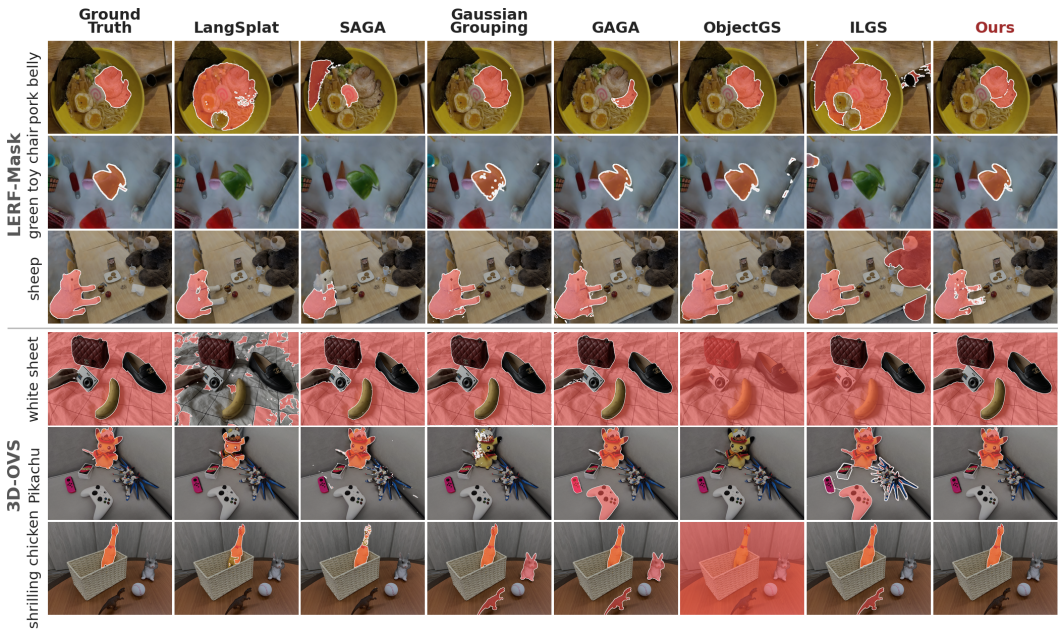


Figure 3: Qualitative comparison on LERF-Mask and 3D-OVS. Our method better isolates queried objects from semantically related or spatially adjacent distractors.

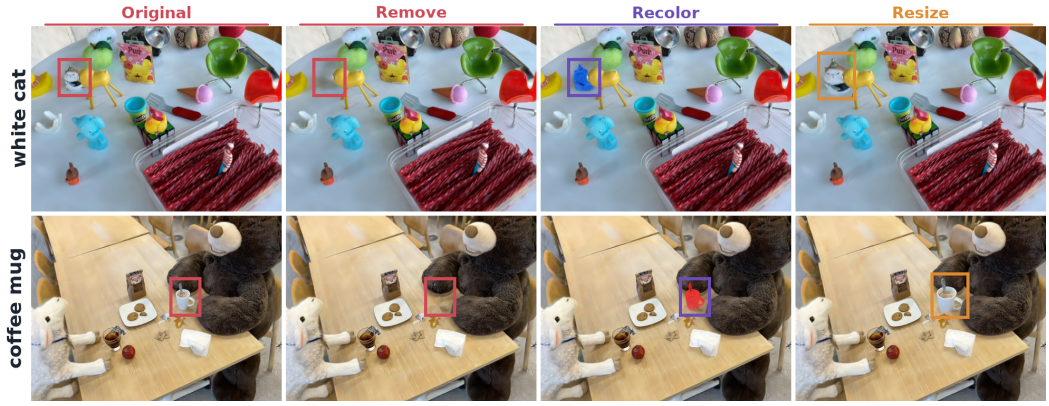


Figure 4: Text-guided 3D editing through scene-global identities.

Efficiency. Table 3 shows that 3D-GIA maintains comparable training efficiency. Under the setup, identity-level querying takes 19 seconds and the complete inference pipeline takes 36 seconds, compared with 68 seconds for Gaussian Grouping and 49 seconds for LangSplat.

Rendering quality. Table 4 shows that geometry regularization slightly improves PSNR and SSIM while keeping LPIPS comparable, indicating that boundary-aware refinement does not sacrifice appearance fidelity for improved segmentation.

Table 3: Efficiency on the *ramen* scene.

Method	Training (min)			Inference (s)		
	Prep.	Opt.	Total	Rend.	Query	Total
Gaussian Grouping	20	50	**70**	∼1	67	68
LangSplat	50	40	90	15	34	49
3D-GIA (**Ours**)	21	54	75	17	19	**36**

Table 4: Novel-view synthesis on LERF-Mask.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
3DGS	27.245	0.891	0.214
GS Grouping (Ye et al., 2024)	27.681	0.897	**0.204**
Ours w/o geometry-aware reg.	27.720	0.903	0.215
Ours	**27.794**	**0.904**	0.213

4.3 Qualitative Results

Figure 3 shows cleaner separation of fine targets and adjacent distractors on both benchmarks.

3D Scene Editing. As shown in Figure 4, scene-global identities directly select Gaussians for object removal, recoloring, and 1.5× isotropic resizing.

4.4 Ablation Study

Table 5 isolates each design choice. Removing the hierarchy module—which replaces hierarchical SAM2 masks with non-hierarchical masks and removes coarse/fine supervision while retaining all other losses— causes the largest drop, by 19.2 mIoU and 20.2 mBIoU. Replacing learned association with DEVA-derived hard identities also reduces performance by 14.0 and 14.3 points, respectively, confirming the value of calibrated, conflict-aware 3D association. Within this branch, identity–semantic transfer contributes 5.9/6.1 points, cannot-links 4.0/3.6, and the identity-pair loss 2.2/2.4; one-hot targets further underperform the soft posterior by 2.4/2.2. Finally, geometry regularization contributes 4.8 mIoU and 4.9 mBIoU, consistent with the rendering gains in Table 4.

Table 5: Ablation study on LERF-Mask (%).

Variant	mIoU ↑	mBIoU ↑
Full model	**88.0**	**82.1**
Scene-global identity modeling		
DEVA-derived identities (Cheng et al., 2023)	74.0	67.8
One-hot association targets	85.6	79.9
w/o cannot-link constraints	84.0	78.5
w/o identity pair loss	85.8	79.7
w/o identity–semantic transfer	82.1	76.0
w/o hierarchy module	68.8	61.9
w/o geometry-aware regularization	83.2	77.2

4.5 Controlled Cross-View Identity Association

Protocol. We project ScanNet ground-truth instances into registered views and independently permute IDs. Both methods receive identical view-local masks, so cross-view identities are used only for evaluation. To remove reconstruction quality, 3D-GIA uses primitives back-projected from ScanNet depth with registered poses, intrinsics, and depth-to-color calibration; DEVA receives the same masks and RGB sequence with native temporal consensus/propagation. We compare calibrated, conflict-aware 3D association with RGB temporal association, plus 3D-overlap-only and no-cannot-link variants.



Figure 5: Cross-view identity association.

Results. Figure 5 shows DEVA changing IDs across views, whereas 3D-GIA preserves scene-global identities. Table 6 reports seven ScanNet scenes: 3D-only uses support overlap and C-link denotes cannot-link. AssA, AQ, IDF1, B^3-F1, and Cov. are percentages; Sw. is a total and Fr. the mean additional IDs per recurring instance. Relative to DEVA, 3D-GIA improves AssA/AQ/IDF1/B^3-F1 by 5.3/7.6/2.3/7.6 points, respectively, while increasing coverage from 92.30 to 98.46 and reducing switches from 848 to 326 (Fr.: 3.03 to 1.15). Removing cannot-links lowers AssA by 4.5 and IDF1 by 6.6 points, showing that explicit conflicts prevent identities from propagating across distinct objects. The 3D-only variant collapses to 39.81 AssA and 46.40 IDF1 despite full coverage; its parenthesized Sw./Fr. values are diagnostic reductions caused by over-merging, rather than improved association.

Table 6: Controlled association on ScanNet.

Method	AssA↑	AQ↑	IDF1↑	B^3-F1↑	Sw.↓	Fr.↓	Cov.↑
DEVA tracking	55.33	49.01	68.51	69.31	848	3.03	92.30
3D-only	39.81	44.26	46.40	59.53	(213)	(0.80)	98.46
w/o C-link	56.20	54.69	64.20	73.45	(277)	(0.96)	98.46
3D-GIA assoc.	**60.65**	**56.58**	**70.83**	**76.94**	**326**	**1.15**	**98.46**

5 Conclusion

3D-GIA converts unordered hierarchical masks into soft scene-global identity posteriors that supervise identity and language; detached field predictions guide geometry. This shared representation enables consistent open-vocabulary segmentation and object-level editing across views. Together, these results establish scene-global identity as a unified interface for multiview association, open-vocabulary querying, and editable 3D representation.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools as writing and formatting assistants. These tools were used to improve the logical clarity and readability of the paper, correct grammatical errors, revise sentence structures for more concise expression, strengthen paragraph coherence, and adjust the manuscript formatting. All AI-assisted revisions were reviewed and verified by the authors. The authors take full responsibility for the final content of the manuscript.

REFERENCES

Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. URL https://aclanthology.org/C98-1012/.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3D gaussians. In *AAAI*, pp. 1971–1979, 2025.

Eunsung Cha, Hyunjoon Lee, and Jaesik Park. Relation-centric open-vocabulary 3D gaussian segmentation. *arXiv preprint arXiv:2607.01140*, 2026.

Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6100–6111, 2025.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander G. Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, et al. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

Antoine Guédon and Vincent Lepetit. SuGaR: Surface-aligned gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In *CVPR*, 2024.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D gaussian splatting for geometrically accurate radiance fields. *ACM Transactions on Graphics*, 43(4), 2024.

Tianyu Huang, Runnan Chen, Dongting Hu, Fengming Huang, Mingming Gong, and Tongliang Liu. OpenInsGaussian: Open-vocabulary instance gaussian segmentation with context-aware cross-view fusion. In *ICCV Workshops*, pp. 6400–6409, 2025.

SungMin Jang and Wonjun Kim. Identity-aware language gaussian splatting for open-vocabulary 3D semantic segmentation. In *ICCV*, pp. 20467–20476, 2025.

Krishna Murthy Jatavallabhula, Hamees Kuwajerwala, Qiao Gu, Muhammad Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso M. de Melo, Kaushik Krishna, Liam Paull, Florian Shkurti, Antonio Torralba, et al. ConceptFusion: Open-set multimodal 3D mapping. In *Robotics: Science and Systems*, 2023.

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 *ICCV*, 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.

Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric P. Xing, and Shijian Lu. Weakly supervised 3D open-vocabulary segmentation. In *NeurIPS*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision–ECCV 2024*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Jonathon Luiten, Aljoša Osep, Patrick Dendorfer, Philip H. S. Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129:548–578, 2021. doi: 10.1007/s11263-020-01375-2.

Weijie Lyu, Xueting Li, Abhijit Kundu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Gaga: Group any gaussians via 3D-aware memory bank. *Transactions on Machine Learning Research*, 2026.

Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Dinh Phung, and Khoi Nguyen. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. In *CVPR*, 2024.

Youngmin Oh, Changjae Oh, and Bumsub Ham. Proto-SaGa: Prototype-based 3D scene segmentation with semantic-aware gaussian grouping. In *CVPR Findings*, pp. 7493–7502, 2026.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, pp. 815–824, 2023.

Yuning Peng, Haiping Wang, Yuan Liu, Chenglu Wen, Zhen Dong, and Bisheng Yang. GAGS: Granularity-aware feature distillation for language gaussian splatting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8376–8384, 2026. doi: 10.1609/aaai.v40i10.37787.

Jens Piekenbrinck, Christian Schmidt, Alexander Hermans, Narunas Vaskevicius, Timm Linder, and Bastian Leibe. OpenSplat3D: Open-vocabulary 3D instance segmentation using gaussian splatting. In *CVPR Workshops*, pp. 5246–5255, 2025.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *CVPR*, 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 of *Proceedings of Machine Learning Research*, 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.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision–ECCV 2016 Workshops*, pp. 17–35, 2016. doi: 10.1007/978-3-319-48881-3_2.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

Hongyu Shen, Junfeng Ni, Yixin Chen, Weishuo Li, Mingtao Pei, and Siyuan Huang. Trace3D: Consistent segmentation lifting via gaussian instance tracing. In *ICCV*, pp. 6656–6666, 2025.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *NeurIPS*, 2023.

Matias Turkulainen, Xiaoming Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. DN-Splatter: Depth and normal priors for gaussian splatting and meshing. In *WACV*, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3D scenes. In *ECCV*, 2024.

Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. OmniSeg3D: Omniversal 3D segmentation via hierarchical contrastive learning. *arXiv preprint arXiv:2407.18206*, 2024.

Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. In *ACM SIGGRAPH Asia Conference Papers*, 2024.

Jiaxin Zhang, Junjun Jiang, Youyu Chen, Kui Jiang, and Xianming Liu. COB-GS: Clear object boundaries in 3DGS segmentation based on boundary-adaptive gaussian splitting. In *CVPR*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Ruijie Zhu, Mulin Yu, Linning Xu, Lihan Jiang, Yixuan Li, Tianzhu Zhang, Jiangmiao Pang, and Bo Dai. ObjectGS: Object-aware scene reconstruction and scene understanding via gaussian splatting. In *ICCV*, 2025.

APPENDIX

This appendix provides additional implementation, evaluation, and reproducibility details for 3D-GIA. Appendix A specifies the association rules, training objectives, fixed hyperparameters, and mask preprocessing settings. Appendix B describes the evaluation metrics, randomness, repeated runs, and computing infrastructure. Appendix C presents the controlled ScanNet association protocol and its sensitivity and statistical analyses. Finally, Appendix D gives the complete optimization and densification schedule.

A REPRODUCIBILITY DETAILS

We provide the support construction, visibility and fragment tests, cannot-link rules, loss normalizations, margins, and fixed hyperparameters omitted from the main paper for space. SAM2 preprocessing uses one shared preset for all LERF-Mask scenes and another shared preset for all 3D-OVS scenes; both are listed explicitly below.

A.1 2D-TO-3D SUPPORT CONSTRUCTION

Let $\boldsymbol{\mu}_i$ denote a Gaussian center after the 8K-iteration 3DGS warm-up. We partition the centers into a regular 3-cm voxel grid. Each nonempty voxel is one support primitive q_r, whose position is the mean of its member centers. This voxelization is used only for offline association; the learned 3DGS still retains its individual Gaussians.

For the standard posed-RGB setting used on LERF-Mask and 3D-OVS, every Gaussian initialized from a COLMAP point stores that point's `point3D_id`, and the ID is inherited by Gaussians created through densification. Let $\mathcal{T}_r$ be the set of valid COLMAP point IDs carried by the Gaussians in primitive q_r. If point $t \in \mathcal{T}_r$ is observed in view v, COLMAP provides its registered image location $\mathbf{x}_{t,v}$, rescaled when needed to the mask resolution. For a mask observation n from view v_n, its 3D support is

$$\mathcal{Q}_n = \{q_r : \exists t \in \mathcal{T}_r \text{ such that } t \text{ is observed in } v_n \text{ and } \mathbf{x}_{t,v_n} \in M_n\}. \tag{11}$$

Thus, a primitive supports a mask only through an actual COLMAP track observation in that view; merely lying inside the camera frustum is insufficient. This construction requires registered RGB images and their COLMAP tracks but no external depth map.

If inherited track IDs are unavailable, the implementation falls back to dense projection: it projects each primitive center with the known camera, retains positive-depth in-image projections inside M_n, and, when a depth map is available, additionally requires the projected depth to agree with the sampled depth within ϵ_d. The controlled ScanNet experiment uses this second path: support points are back-projected from the official depth stream, pooled into the same 3-cm voxels, and filtered by registered color-view depth consistency. Its stride, depth range, and ϵ_d are specified in Appendix C. In either path, the result stored for observation n is the finite primitive set $\mathcal{Q}_n$; all cross-view association cues below are computed from these sets and their camera projections.

A.2 VISIBILITY AND BIDIRECTIONAL REPROJECTION

Let $\mathcal{Q}_a$ be the voxelized 3D support of observation a, and let M_b be the foreground-pixel set of observation b in its view. Their support overlap is

$$a_{ab}^{3D} = \max\left(\frac{|\mathcal{Q}_a \cap \mathcal{Q}_b|}{|\mathcal{Q}_a|}, \frac{|\mathcal{Q}_a \cap \mathcal{Q}_b|}{|\mathcal{Q}_b|}\right). \tag{12}$$

When projecting it into view b, a primitive is valid if it has positive camera depth and its projected center lies inside the image. Let $\mathcal{P}_{a\to b}$ be the set of distinct valid projected pixels. We define

$$\begin{aligned} v_{a\to b} &= \frac{\#\{q \in \mathcal{Q}_a : q \text{ projects validly into } b\}}{|\mathcal{Q}_a|}, \\ r_{a\to b} &= \frac{|\mathcal{P}_{a\to b} \cap M_b|}{|\mathcal{P}_{a\to b} \cup M_b|}. \end{aligned} \tag{13}$$

The bidirectional reprojection evidence is $a_{ab}^{rep} = \frac{1}{2}(r_{a\to b} + r_{b\to a})$. The projection is considered sufficiently visible when $\max(v_{a\to b}, v_{b\to a}) > 0.50$.

A.3 CANNOT-LINKS AND VERIFIED FRAGMENTS

The implementation first constructs a provisional conflict $\widetilde{c}_{ab}$ when either

1. the pair comes from the same view, has mask IoU below 0.05, and has symmetric containment at most 0.85; or
2. $\max(v_{a\to b}, v_{b\to a}) > 0.50$ but $\min(r_{a\to b}, r_{b\to a}) < 0.10$.

The optional support-distance test is disabled in the reported setting (threshold 0). Let z_{ab}^{frag} and z_{ab}^{par} denote acceptance by the verified-fragment and verified-same-parent tests below, and let c_{ab}^{hier} and c_{ab}^{par-} denote a direct child–parent residual conflict and contradictory verified parents. The final edge flag is

$$c_{ab} = \left[\widetilde{c}_{ab} \wedge \neg(z_{ab}^{frag} \vee z_{ab}^{par})\right] \vee c_{ab}^{hier} \vee c_{ab}^{par-}. \tag{14}$$

Thus a verified same-view relation clears only the provisional geometric conflict; hierarchy and contradictory-parent conflicts remain final edge-level cannot-links. A concept conflict, when enabled, behaves in the same way, but it is disabled in the reported experiments.

Two same-view atomic masks are treated as verified fragments only if their IoU is at most 0.01, their containment is at most 0.35, their appearance cosine similarity is at least 0.70, and the minimum distance between their 3D supports is at most 0.07 m. They must additionally share the same coarse parent with parent overlap at least 0.65. At most two fragments from one view may occur in a cluster. Their fragment score is

$$a_{ab}^{frag} = \tfrac{1}{2} a_{ab}^{app} + \tfrac{1}{2}\left[1 - \frac{d(\mathcal{Q}_a, \mathcal{Q}_b)}{0.07}\right]_+. \tag{15}$$

Here a_{ab}^{app} is appearance cosine similarity, and $d(\mathcal{Q}_a, \mathcal{Q}_b)$ is the minimum Euclidean distance between their primitive centers.

The same-parent verifier uses the same IoU, containment, parent-identity, and parent-overlap tests, appearance similarity at least 0.55, and support distance at most 0.06 m. It supplies $a_{ab}^{par} > 0$ and obeys the same two-observations-per-view cap. Either verified relation can clear provisional same-view disjointness, but neither can clear a hierarchy or parent conflict. Its score is

$$\begin{aligned} a_{ab}^{par} = 0.45 a_{ab}^{app} + 0.35 \left[1 - \frac{d(\mathcal{Q}_a, \mathcal{Q}_b)}{0.06}\right]_+ \\ + 0.20 \min(o_a^{par}, o_b^{par}), \end{aligned} \tag{16}$$

where o_a^{par} is the overlap of observation a with its verified parent.

For an admissible pair, the continuous edge score is

$$\begin{aligned} e_{ab} = 0.55 a_{ab}^{3D} + 0.25 a_{ab}^{rep} + 0.35 a_{ab}^{app} \\ + 0.20 a_{ab}^{frag} + 0.30 a_{ab}^{par}, \end{aligned} \tag{17}$$

where an inapplicable verified-relation term is zero. The coefficients are not normalized because this score ranks pairs that pass the discrete gate; a final cannot-link always vetoes a merge.

Initial unions require the final $c_{ab} = 0$ and either (i) appearance above 0.42 together with 3D support overlap above 0.32 or reprojection evidence above 0.06, or (ii) a verified fragment/same-parent relation. Ordinary pairs must be view-compatible; verified same-view relations may place at most two observations from one view in a component. In other words, appearance is the semantic gate and the two geometric cues are alternatives rather than three simultaneous requirements. The continuous cue combination orders pairs that pass this gate; it cannot override a final cannot-link. In the standard cluster-pair pass, pairs containing any final cannot-link are discarded; otherwise, the largest eight cross-cluster edge scores are averaged and the clusters are merged when this mean exceeds 0.46.

A.4 GAUSSIAN-LEVEL SOFT IDENTITY POSTERIOR

Let q_n be the product of SAM2's predicted-IoU and stability scores for observation n. Its lifting confidence is

$$b_n = q_n \begin{cases} \min(1, |\mathcal{Q}_n|/12), & \text{COLMAP-track support,} \\ d_n \min(1, |\mathcal{Q}_n|/N_n), & \text{dense projection,} \end{cases} \tag{18}$$

where N_n is the number of valid projected samples inside the mask and d_n is their depth-consistency fraction ($d_n = 1$ when depth is unavailable). For cluster k, the effective vote is

$$\beta_n = b_n \left[0.20 + 0.80 \min(V_k/12, 1)\sqrt{\bar{b}_k}\right], \tag{19}$$

where V_k and $\bar{b}_k$ are its number of supporting views and mean lifting confidence. For every primitive, these votes are accumulated per global cluster and normalized to obtain $\mathbf{s}_r$; its concentration γ_r is the largest normalized cluster mass. Nearest-center transfer assigns $(\mathbf{s}_i, \gamma_i)$ to Gaussian i. Gaussians with $\gamma_i \le \tau_{id}$ are excluded from the soft cross-entropy; we set $\tau_{id} = 0.55$. Before normalization, we retain at most the two strongest cluster scores whose value is at least 0.25 of the maximum. We use at most 32,768 Gaussians and 4,096 positive/negative pairs per update. The identity loss uses

$$(\lambda_{soft}, \lambda_{pair}, m_{id}) = (0.10, 0.05, 0.30). \tag{20}$$

A.5 Hierarchy loss normalization and margins

For a region Ω of area A_Ω, its inverse-area weight is

$$w_\Omega = \mathrm{clip}\left[\left(\frac{\mathrm{median}_{\Omega'} A_{\Omega'}}{A_\Omega}\right)^{0.45}, 0.5, 5.0\right]. \tag{21}$$

Overlapping regions use the maximum applicable weight. For $g \in \{f, c\}$, the normalized weighted cross-entropy is

$$\mathcal{L}_g = \frac{1}{\log 256} \frac{\sum_{p\in\mathcal{R}_g} w_p \,\mathrm{CE}(\mathbf{l}_p, y_p)}{\sum_{p\in\mathcal{R}_g} w_p}. \tag{22}$$

Here $\mathcal{R}_g$ is the set of valid pixels at level g, $\mathbf{l}_p$ and $P_p = \mathrm{softmax}(\mathbf{l}_p)$ are the logits and probabilities, y_p is the target, and w_p is the maximum weight of regions covering p. We sample at most 4,096 valid pixels per hierarchy term. With $\delta_h = 0.12$, the child and parent margins are

$$\mathcal{M}_f = \mathbb{E}_{p\in\Omega_{child}}[P_p(y_{parent}) - P_p(y_{child}) + \delta_h]_+, \tag{23}$$

$$\mathcal{M}_c = \mathbb{E}_{p\in\Omega_{parent\setminus children}}\Big[\max_{j\in\mathrm{child}(y_{parent})} P_p(j) - P_p(y_{parent}) + \delta_h\Big]_+. \tag{24}$$

The four coefficients are $(\lambda_f, \lambda_c, \lambda_f^m, \lambda_c^m) = (0.18, 0.12, 0.04, 0.03)$.

The dense object cross-entropy uses inverse-area power 0.5, weights clipped to $[0.25, 8.0]$, and division by $\log 256$. Its coefficient is $\lambda_{obj} = 1.0$. The language reconstruction coefficient is $\lambda_{lang} = 1.0$.

A.6 Semantic transfer and geometry weights

For completeness, the identity–semantic transfer summarized in the main paper uses the detached teacher. Following the main-paper notation, $\mathbf{s}_i^+$ is the association posterior zero-padded to the fixed classifier label space:

$$\begin{aligned} \mathbf{r}_i &= \mathrm{Norm}_1\big(\mathbf{s}_i^+ \odot \mathrm{sg}(\mathbf{p}_i)\big), \\ w_i &= \mathrm{sg}\big(\gamma_i \|\mathbf{p}_i\|_\infty (\mathbf{s}_i^+)^\top \mathbf{p}_i\big). \end{aligned} \tag{25}$$

More precisely, $\mathcal{V}_{\mathrm{sem}} = \{i : \gamma_i > 0.65, \|\mathbf{p}_i\|_\infty > 0.45, (\mathbf{s}_i^+)^\top \mathbf{p}_i > 0\}$ is the reliable Gaussian set, and $\mathcal{K}^+ = \{k : \sum_{i\in\mathcal{V}_{\mathrm{sem}}} \gamma_i r_{i,k} \ge 8\}$ is the supported identity set. After restricting and renormalizing $\mathbf{r}_i$ over $\mathcal{K}^+$ to obtain $\widehat{\mathbf{r}}_i$, the detached prototype and per-Gaussian target are

$$\boldsymbol{\pi}_k^{\mathrm{sem}} = \mathrm{sg}\left[\mathrm{Norm}_2\left(\sum_{i\in\mathcal{V}_{\mathrm{sem}}} w_i r_{i,k} \bar{\mathbf{z}}_i\right)\right], \tag{26}$$

$$\widetilde{\boldsymbol{\pi}}_i^{\mathrm{sem}} = \mathrm{Norm}_2\left(\sum_{k\in\mathcal{K}^+} \widehat{r}_{i,k} \boldsymbol{\pi}_k^{\mathrm{sem}}\right). \tag{27}$$

Let $\widehat{w}_i = w_i / \sum_{j \in \mathcal{V}_{\text{sem}}} w_j$, $\mathbf{M} = [\boldsymbol{\pi}_k^{\text{sem}}]_{k \in \mathcal{K}^+}$, and $\mathbf{q}_i = \text{softmax}(\mathbf{M}^\top \bar{\mathbf{z}}_i / T)$. The exact objective is

$$\begin{aligned}\mathcal{L}_{id\text{-}sem} = & \sum_{i \in \mathcal{V}_{\text{sem}}} \widehat{w}_i \Big[\lambda_{cmp}(1 - \bar{\mathbf{z}}_i^\top \widetilde{\boldsymbol{\pi}}_i^{\text{sem}}) + \lambda_{ce}\, \text{CE}(\widehat{\mathbf{r}}_i, \mathbf{q}_i) \Big] \\ & + \lambda_{sp} \ell_{pair}(\bar{\mathbf{z}}; \mathcal{P}^{\pm}_{\mathcal{V}_{\text{sem}}}, m_{sem}),\end{aligned} \tag{28}$$

where ℓ_{pair} is the cosine positive/cannot-link loss defined in the main paper.

The pair sets $\mathcal{P}^{\pm}_{\mathcal{V}_{\text{sem}}}$ restrict the main-paper positive and cannot-link pairs to endpoints in $\mathcal{V}_{\text{sem}}$.

Identity-conditioned semantic transfer begins at iteration 14K. We use temperature $T = 0.10$, posterior-concentration and identity-prediction confidence thresholds 0.65 and 0.45, minimum teacher mass 8, semantic pair margin 0.25, and

$$(\lambda_{cmp}, \lambda_{ce}, \lambda_{sp}) = (0.030, 0.0125, 0.010). \tag{29}$$

The object-internal part regularizer begins at 16K and is evaluated every 30 iterations. Within each supervised global-identity region, it aligns each 2D language segment's mean code with its frozen segment feature and imposes a cosine-similarity margin of 0.65 between different segments. Its compactness and separation weights are 0.04 and 0.01, respectively. Geometry regularization begins at 16K. The flattening coefficient is 0.06; the separation and continuity coefficients are 0.15 and 0.20, respectively. Let R denote the scene camera extent. The scale and footprint thresholds are $\tau_s = 0.05R$ and $\tau_a = \tau_s^2$, both with unit penalty weights. For the surface affinity, $\sigma_n = 0.25$; σ_d and σ_x are respectively the median point-to-plane deviation and median squared neighbor distance in the current sampled neighborhood. We use $m_{\text{geom}} = 0.5$. Neighborhoods use $k = 20$, a 40° normal gate, and 16 neighbors for the local PCA normal. The Gaussian-normal mixing coefficient is linearly increased from 0.1 to 0.5 during geometry warm-up. Each update retains at most 100,000 Gaussians, samples 1,000 anchors, and is evaluated every 12 iterations.

A.7 Fixed inference parameters

For LERF-Mask, a Grounded-SAM query mask from the first evaluation view selects every identity region whose intersection-over-region-area exceeds 0.70. The selected scene-global identities are then fixed for all views, and their summed posterior is thresholded at 0.30.

For 3D-OVS identity-level text querying, the similarity score averages the highest 70% of pixels in each identity region. We retain at most eight identities whose score reaches 85% of the best identity score and threshold their summed posterior at 0.30. Thus

$$(\kappa, \tau_{rel}, \tau_m) = (0.70, 0.85, 0.30). \tag{30}$$

This triple is selected once on held-out training views and then fixed for all reported 3D-OVS scenes. When a hierarchy mapping is supplied, a selected parent is recursively expanded to its descendants before summing the probabilities.

A.8 Hyperparameter development protocol

Table 7 summarizes the principal values evaluated during method development. The SAM2 presets are reported separately in Table 8: the LERF-Mask setting was developed on *figurines* training views and then fixed for the other two scenes; the 3D-OVS setting was fixed using training views across its five scenes. Four complete proposal templates were evaluated first, followed by component-level refinement. All hyperparameter and preprocessing-preset selection uses a fixed held-out subset of registered training views, shared by every candidate. Training-side annotations are masks attached only to these training views and are disjoint from every view and mask used for the reported metrics; they are used for configuration ranking, never for model optimization. Candidates are ranked by macro mIoU on this fixed subset, with boundary quality, cross-view association diagnostics, and RGB reconstruction used only to resolve close settings. Evaluation views, evaluation annotations, and reported test metrics never enter selection. All parameters not shown in this table were inherited from the base 3DGS implementation or fixed once; their final values are given throughout this document.

Table 7: Development search summary and selected values. Explicit sets list all candidates for the principal association thresholds; $n : [a, b]$ denotes n discrete candidates spanning endpoints a and b for grouped loss weights. Zero-valued candidates are module ablations.

Parameter group	Development candidates	Selected
SAM2 preprocessing	Four proposal templates followed by component-level refinement	Table 8
Association thresholds	$\tau_{3D} : \{.30, .32, .35, .40\}$; $\tau_{rep} : \{.05, .06, .08\}$; $\tau_{app} : \{.36, .42, .45\}$; $\tau_C : \{.42, .46, .50, .55\}$	$(.32, .06, .42, .46)$
Soft targets	Concentration: $\{.45, .55\}$; $\lambda_{soft} : 7 : [0, .50]$	$.55$; $.10$
Identity pairing	$\lambda_{pair} : 5 : [0, .10]$; margin: $\{.30, .32\}$	$.05$; $.30$
Hierarchy	Margin: $\{.12, .14, .15\}$; child/parent CE: $6 : [0, .475777]/7 : [0, .45528]$; child/parent margin: $5 : [0, .120]/5 : [0, .147]$	$.12$; $.18/.12$; $.04/.03$ $.10$;
Semantic mapping	Temperature: $\{.08, .10\}$; confidence: $\{.55, .60, .65\}/\{.35, .40, .45\}$; $(\lambda_{cmp}, \lambda_{ce}, \lambda_{sp}) : 5 : [0, .08]/7 : [0, .04]/5 : [0, .03]$	$.65/.45$; $.03/.0125/.01$
Geometry regularization	$\lambda_{flat} : \{0, .05, .06\}$; $\lambda_{sep} : \{0, .03, .08, .15\}$; $\lambda_{cont} : \{0, .05, .10, .12, .20\}$	$(.06, .15, .20)$

A.9 Final mask-preprocessing parameters

SAM2.1 Hiera-L uses 64 points per batch and crop point downscale factor two. Table 8 lists each shared coarse/fine preset in the order (points per side, predicted-IoU threshold, stability threshold, NMS threshold, crop layers, minimum region area, minimum area, maximum area ratio, maximum masks). The presets are fixed before evaluation without using evaluation annotations or metrics. All three LERF-Mask scenes use the first row, and all five 3D-OVS scenes use the second.

Table 8: Final hierarchical SAM2 proposal parameters.

Scene(s)	Coarse SAM2 tuple	Fine SAM2 tuple
LERF-Mask (all)	$(28, .84, .93, .58, 0, 180, 420, .90, 70)$	$(32, .80, .91, .74, 1, 96, 120, .985, 140)$
3D-OVS (all)	$(32, .82, .91, .64, 0, 120, 220, .94, 90)$	$(40, .78, .89, .76, 1, 64, 70, .985, 170)$

The first six tuple entries are arguments of SAM2's official automatic mask generator; their values are our fixed benchmark presets rather than SAM2 defaults. The final three entries are our post-generation area and count filters.

Coarse/fine fusion uses duplicate IoU 0.85, parent containment 0.80, maximum coarse-duplicate IoU 0.55, maximum coarse-duplicate containment 0.95, and thin-mask aspect ratio 4.0. The tuple (minimum fine area, maximum fine area ratio, maximum child/parent area ratio, maximum masks) is $(20, 0.18, 0.35, 180)$ for LERF-Mask and $(20, 0.20, 0.40, 180)$ for 3D-OVS.



Figure 6: Hierarchical mask preprocessing on the *ramen* scene using the shared LERF-Mask preset. Coarse proposals capture object extents, fine proposals recover parts and thin structures, and fusion records parent–child relations. Parent residuals and retained fine masks are subsequently lifted as separate association observations.

B Evaluation and Reporting Protocol

B.1 Evaluation metrics

For a binary prediction P and ground-truth mask G, intersection over union is

$$\mathrm{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}. \tag{31}$$

We first average over annotated queries/classes and then over scenes to obtain mIoU, so frequent or large objects do not dominate the score. Boundary IoU (Cheng et al., 2021) applies the same expression to morphological boundary bands ∂P and ∂G. Following the released LERF-Mask evaluator, the band width is 2% of the image diagonal. We report mBIoU because it is more sensitive than region IoU to leakage across adjacent object boundaries.

For controlled association, let n_{gp} be the number of common observations assigned to ground-truth identity g and predicted identity p, with $n_g = \sum_p n_{gp}$, $n_p = \sum_g n_{gp}$, and $N = \sum_{g,p} n_{gp}$. Following HOTA's association construction (Luiten et al., 2021), the reported common-observation association accuracy is

$$\mathrm{AssA} = \frac{1}{N} \sum_{g,p} n_{gp} \frac{n_{gp}}{n_g + n_p - n_{gp}}, \tag{32}$$

where each summand is repeated for the n_{gp} observations in its cell. Let $\mathcal{G}_{\mathrm{gt}}$ be the set of ground-truth identities. The ground-truth-balanced association quality is

$$\begin{aligned} \mathrm{AQ} = \frac{1}{|\mathcal{G}_{\mathrm{gt}}|} \sum_g \frac{1}{n_g} \\ \sum_p \frac{n_{gp}^2}{n_g + n_p - n_{gp}}. \end{aligned} \tag{33}$$

After Hungarian matching, IDF1 (Ristani et al., 2016) is the matched-observation fraction. B^3 (Bagga & Baldwin, 1998) precision and recall are

$$\begin{aligned} P_{B^3} &= \frac{1}{N} \sum_{g,p} \frac{n_{gp}^2}{n_p}, \\ R_{B^3} &= \frac{1}{N} \sum_{g,p} \frac{n_{gp}^2}{n_g}, \end{aligned} \tag{34}$$

and B^3-F1 is their harmonic mean. Identity switches count changes of the predicted identity between consecutive observations of one ground-truth instance; fragments count the number of additional predicted identities per ground-truth instance; coverage is the fraction of annotated instance pixels assigned a non-background prediction. These complementary quantities expose fragmentation and over-merging that a single overlap score can hide.

For RGB rendering, PSNR measures pixel fidelity, SSIM (Wang et al., 2004) measures local structural similarity, and LPIPS (Zhang et al., 2018) measures learned perceptual distance. They are included to check that segmentation-oriented geometry optimization does not sacrifice novel-view appearance quality.

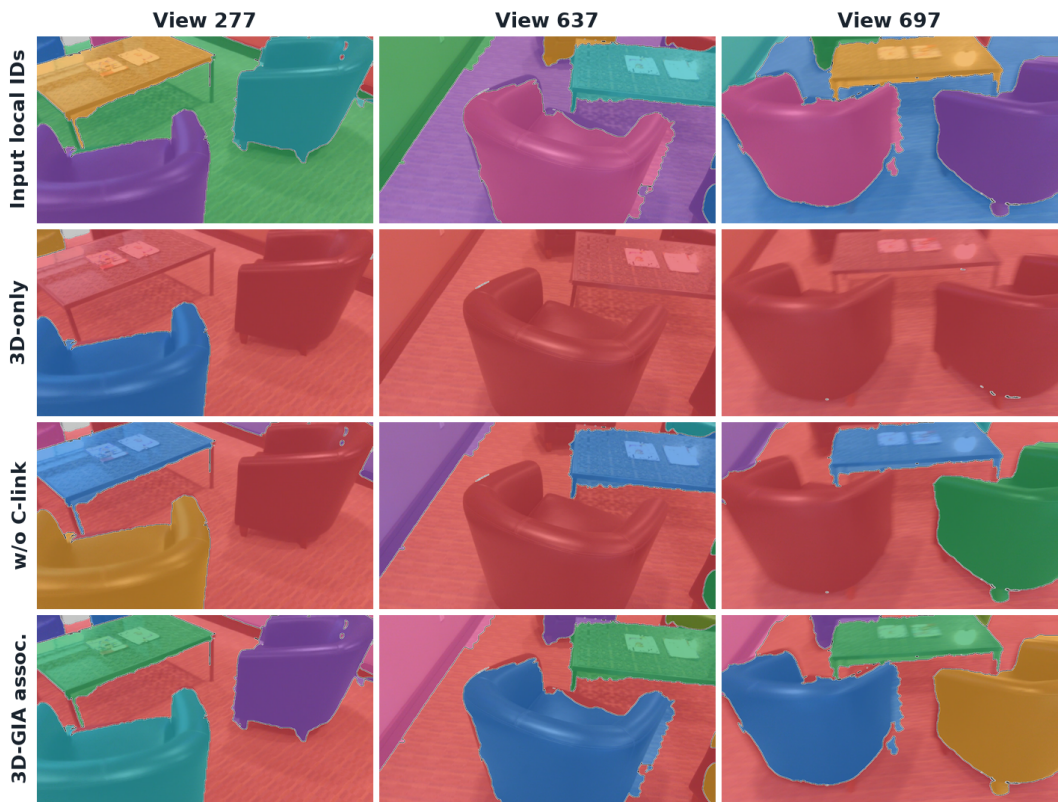


Figure 7: Controlled association on ScanNet scene `scene0192_00`. The top row shows independently permuted view-local identities. Support overlap alone over-merges most instances; removing cannot-links leaves several distinct regions merged, whereas full 3D-GIA association preserves separate global identities across views.

B.2 Randomness and number of runs

The main-paper LERF-Mask and 3D-OVS tables report the complete seed-0 run for each scene. For LERF-Mask, we additionally repeat full 3D-GIA training with seeds 1 and 2 and report all three runs in Table 9; these repeats do not replace or alter the submitted main-paper values. LERF-Mask ablations use seed 0. Each training seed is applied to Python, NumPy, PyTorch, and all visible CUDA random generators. Preprocessing and 3D-GIA association are deterministic for fixed inputs. The ScanNet protocol uses seed 20260722 only to permute the ground-truth identity numbers independently in each view. CUDA reductions can cause small machine-dependent floating-point differences, so bitwise equality across hardware and driver versions is not assumed.

Table 9: Run-to-run variation of full 3D-GIA on LERF-Mask (%). Each entry is the macro average over the three scenes under the main-paper evaluation protocol.

Training seed	mIoU	mBIoU
0 (main paper)	88.00	82.10
1 (additional)	87.75	81.30
2 (additional)	86.26	80.77
Mean $\pm$ SD	87.34 ± 0.94	81.39 ± 0.67

The controlled ScanNet result aggregates one deterministic association run on each of seven scenes. Baseline values without a reproduction marker are taken from the cited papers; reproduced baselines use their released code.

B.3 COMPUTING INFRASTRUCTURE

Experiments were run on Ubuntu 20.04.6 LTS with an Intel Xeon Platinum 8458P host, 512 GB system memory, and one NVIDIA RTX A6000 with 48 GB memory per scene. The NVIDIA driver was 550.78. The environment used Python 3.8.20, PyTorch 2.1.0, CUDA 12.1, cuDNN 8.9.2, NumPy 1.24.3, SciPy 1.10.0, scikit-learn 1.3.2, OpenCV 4.13.0, Pillow 10.4.0, SAM2.1 Hiera-L, and OpenCLIP 2.32.0 with LAION-2B ViT-B/16 weights. SAM2 preprocessing used a separate Python 3.10.20 environment with PyTorch 2.5.1+cu121, torchvision 0.20.1+cu121, and NumPy 2.2.6. The SAM2 revision is `2b90b9f5ceec907a1c18123530e92e794ad901a4`. All Gaussian Grouping, LangSplat, and 3D-GIA timings in the main-paper efficiency table were remeasured on the same RTX A6000. There, Prep. denotes method-specific offline supervision construction, including 3D-GIA association, and Opt. denotes complete field optimization. The reproduction marker used in result tables does not apply to these unified-hardware timings.

C CONTROLLED SCANNET ASSOCIATION PROTOCOL

We use the same seven-scene subset as Gaga:

`scene0010_00`, `scene0012_00`, `scene0033_00`, `scene0038_00`, `scene0088_00`, `scene0113_00`, and `scene0192_00`.

The subset contains 639 registered evaluation views and 189 ground-truth instances visible in at least two views. Unannotated pixels are ignored.

Both methods receive identical projected ground-truth instance masks, while the mask IDs are independently permuted in every view. Ground-truth cross-view identities are withheld until evaluation. 3D-GIA constructs support primitives from ScanNet-provided depth maps, camera poses, intrinsics, and depth-to-color calibration; DEVA tracking receives the corresponding RGB frames and the same mask observations. We use a 3-cm voxel size, depth consistency tolerance 0.08 m, depth stride four, maximum depth 6 m, minimum mask area 64 pixels, and at least one support primitive. The four principal merge thresholds remain $(0.32, 0.06, 0.42, 0.46)$. Same-view fragments use the same $(0.01, 0.35, 0.70, 0.07\ \text{m})$ tests, with at most three fragments per view because the oracle masks contain no object–part parent hierarchy. Accordingly, their edge score uses $0.25a_{ab}^{frag}$ and omits the parent term. All settings are fixed for the seven scenes.

C.1 THRESHOLD SENSITIVITY

We vary one principal association threshold at a time while fixing the other three. This diagnostic uses the controlled ScanNet protocol; the paper settings (bold) were fixed beforehand using the development criterion in Table 7.

Table 10: One-at-a-time association-threshold sensitivity on the seven controlled ScanNet scenes.

Threshold	Value / pooled B^3-F1 (%)			
τ_{3D}	.30/76.94	**.32/76.94**	.35/76.87	.40/76.46
τ_{rep}	.05/76.94	**.06/76.94**	.08/76.94	–
τ_{app}	.36/76.73	**.42/76.94**	.45/77.20	–
τ_C	.42/76.94	**.46/76.94**	.50/76.94	.55/76.94

Across the tested neighborhoods, the change in pooled B^3-F1 from the fixed setting ranges from -0.48 to $+0.26$ percentage points; τ_{rep} is invariant over its tested range.

C.2 SCENE VARIATION AND STATISTICAL TESTS

We treat a scene, rather than an observation or pixel, as the independent unit. Table 11 reports the sample standard deviation across seven scenes and a 95% percentile-bootstrap confidence interval for the scene mean using 10,000 resamples and seed 20260730.

Table 11: Pooled and scene-level B^3-F1 (%) on controlled ScanNet.

Method	Pooled	Scene mean ± SD	95% CI
DEVA tracking	69.31	71.48 ± 15.36	$[61.16, 82.21]$
3D-only	59.53	59.03 ± 22.39	$[44.69, 75.17]$
w/o C-link	73.45	75.70 ± 11.28	$[68.24, 83.53]$
3D-GIA	**76.94**	$\mathbf{79.25 \pm 12.02}$	$[\mathbf{71.22}, \mathbf{87.63}]$

For the two prespecified directional alternatives (3D-GIA higher), one-sided exact paired Wilcoxon signed-rank tests give $W = 27$ against DEVA tracking and $W = 28$ against 3D-only. Both Holm-adjusted p-values are 0.0156.

D Optimization Schedule

Gaussian geometry, opacity, RGB appearance, and language codes are jointly warmed up for 8K iterations using $\mathcal{L}_{rgb}$ and $\mathcal{L}_{lang}$. Offline association is then computed once from the warm-up Gaussians. $\mathcal{L}_{obj}$ is enabled immediately afterward; $\mathcal{L}_{3D\text{-}GIA}$ and $\mathcal{L}_{hier}$ start at 12K, $\mathcal{L}_{id\text{-}sem}$ at 14K, and the object-internal part regularizer and $\mathcal{L}_{geom\text{-}reg}$ at 16K. The identity-guided geometry term is linearly warmed from 16K to 20K. Training ends at 24K iterations. Each scene uses one NVIDIA RTX A6000. The identity classifier has 256 outputs for LERF-Mask and 3D-OVS; unused outputs receive zero target mass.

D.1 Base optimizer and densification parameters

The spherical-harmonic degree is 3. Position learning rate decays from 1.6×10^{-4} to 1.6×10^{-6} with delay multiplier 0.01; feature, language-feature, semantic-feature, opacity, scale, and rotation learning rates are respectively 2.5×10^{-3}, 2.5×10^{-3}, 1.0×10^{-3}, 0.05, 0.005, and 0.001. The DSSIM mixture coefficient is 0.2 and the dense fraction is 0.01. Densification starts at iteration 500, runs every 100 iterations until iteration 10K with gradient threshold 2×10^{-4}, and opacity is reset every 3K iterations. All remaining method-specific parameters are given in the preceding sections.