TOFUGS: TOPOLOGY BEFORE FUSION FOR OPEN-VOCABULARY GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Compact open-vocabulary 3D representations must consolidate multiview, multi-scale semantics into one deployable descriptor per primitive. A common Gaussian splatting pipeline fuses these observations before semantic refinement, implicitly treating fusion as an independent step of refinement. However, this ordering matters because mask granularity can change feature directions as well as the induced refinement topology: fusion before topology constructs a single topology from mixed features and removes branch-specific information, whereas topology before fusion preserves the complementary partitions and mask registrations induced by each branch. To address this problem, we introduce ToFuGS (Topology-before-Fusion Gaussian Splatting), which independently refines branch-specific mask-derived fields first and uses training-view evidence to fuse them into a single deployable descriptor per Gaussian. On LERF, ToFuGS improves its underlying refinement pipeline by 7.34 mIoU points and establishes a new state of the art. Without retuning, it improves results on 3D-OVS, and an order control with a second refinement backend likewise favors refine-then-fuse.

1 INTRODUCTION

Open-vocabulary 3D scene representations make reconstructed scenes searchable with natural language by attaching vision-language features to spatial primitives. LERF introduces multiscale language queries in a NeRF (Kerr et al., 2023), while LangSplat (Qin et al., 2024), LEGaussians (Shi et al., 2024), and Feature 3DGS (Zhou et al., 2024) build on the explicit, efficiently rendered primitives of 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023). Constructing a compact field with one semantic descriptor per Gaussian requires consolidating observations across views and mask granularities.

SFS (Xiong et al., 2026) combines mask-derived descriptors from different granularities before clustering and mask-based semantic refinement. Segment Anything (SAM) supplies masks across these scales (Kirillov et al., 2023), and CLIP maps their visual content into a shared embedding space (Radford et al., 2021).

Consider a toy whose leg is isolated by one mask and grouped with the torso by another. The two descriptors can induce different clusters and mask registrations for the same Gaussians. If averaging merges the leg and torso into one cluster, subsequent refinement propagates semantic updates to both regions.

Because refinement groups are constructed from the feature field, fusing the features first can also change the groups on which refinement operates. Therefore, fusion order affects not only the input features but also the refinement structure.

We call these cluster assignments and mask registrations a *refinement topology*. ToFuGS[1] first refines a cross-scale consensus branch, Uniform, and an intermediate-scale branch, Mid, under their separately induced topologies, and then fuses the refined fields using training-view evidence. Mid retains localized structures that cross-scale averaging can attenuate (Figure 1).

[1]Anonymous code repository: `https://anonymous.4open.science/r/ToFuGS-E146`.

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.

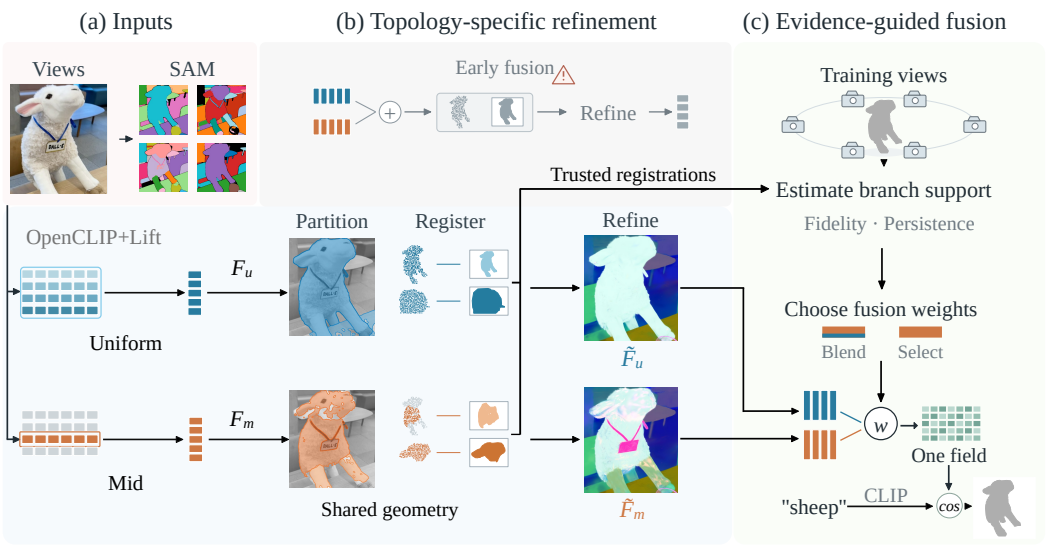


Figure 1: Overview of ToFuGS. (a) Multiscale SAM masks and OpenCLIP descriptors form Uniform (cross-scale consensus, blue) and Mid (orange) fields. (b) Each field is refined using its own Gaussian partition and mask registrations. (c) Training-view evidence guides blending or branch selection, producing one descriptor per Gaussian for text-query evaluation. Figure 13 details support estimation and gating.

Our contributions are:

- We identify fusion order as a structural design choice in semantic refinement.
- We introduce ToFuGS, which preserves branch-specific refinement topologies before evidence-guided fusion into a single deployable Gaussian field.
- We show that refine-then-fuse outperforms fuse-then-refine with two refinement backends, and that the same configuration transfers to 3D-OVS.

2 Related Work

Semantic fusion in open-vocabulary 3D fields. Recent methods make different choices about what semantic evidence to combine and how to encode or render it. LERF integrates multiscale CLIP features in a continuous radiance field (Kerr et al., 2023); Feature 3DGS distills general feature channels into explicit Gaussians (Zhou et al., 2024); LangSplat learns a compact latent from hierarchical SAM semantics (Qin et al., 2024); and LEGaussians couples feature quantization with multiview smoothing (Shi et al., 2024). LangSplatV2 represents each Gaussian through sparse coefficients over a global feature dictionary, allowing high-dimensional features to be rendered efficiently (Li et al., 2025). Dr. Splat assigns CLIP embeddings to the dominant Gaussians along each pixel ray and compresses the resulting descriptors with product quantization (Kim et al., 2025). Other designs move fusion into the renderer: 3D Vision-Language Gaussian Splatting combines cross-modal rasterization with view blending (Peng et al., 2025). This leaves open where fusion should occur relative to feature-dependent refinement. We study this ordering directly.

Semantic features as structure. Feature-dependent grouping makes fusion order a separate design variable. Gaussian Grouping learns identity structure from SAM masks with 3D consistency regularization (Ye et al., 2024). LaGa learns affinity features to associate multiview masks with object-level Gaussian groups, then clusters each object's mask descriptors into a set of semantic representatives (Cen et al., 2025). At query time, it aggregates their relevance scores using descriptor weights. Grouping thus establishes the objects over which multiview semantics are organized. SFS lifts features through a regularized sparse inverse problem and uses the lifted semantic features themselves for clustering (Xiong et al., 2026). The resulting groups filter mask observations before

the field is updated. Our formulation makes these cluster assignments and registrations explicit so that they can be constructed separately for each input field.

Training-view evidence. Multiview consistency has also been used to assess semantic or geometric confidence. LoRA3D, for example, reweights geometric confidence before generating pseudo-labels for scene adaptation (Lu et al., 2025). ToFuGS uses cross-view agreement among registered mask descriptors to weight the two refined fields at each Gaussian.

3 METHOD

3.1 PROBLEM SETTING AND OVERVIEW

Let a reconstructed scene contain N Gaussians and let $\mathcal{V}$ denote its training views. Each view provides SAM masks and their D-dimensional OpenCLIP descriptors, with $D = 512$. Let $\mathcal{Q} = \{0, 1, 2, 3\}$ index the default, small, medium, and large SAM mask channels. For $q \in \mathcal{Q}$, a regularized sparse inverse problem lifts the corresponding observations to a Gaussian feature field $L_q \in \mathbb{R}^{N \times D}$ (Xiong et al., 2026). We use $\mathrm{norm}(x) = x / \max(\|x\|_2, \epsilon)$ row-wise throughout.

For a field F, let $c_F(g)$ denote the cluster label of Gaussian g and $\mathcal{A}_F$ the mask–cluster registrations across training views. We call $\mathcal{T}(F) = (c_F, \mathcal{A}_F)$ a *refinement topology*, and write $\mathcal{R}_{\mathcal{T}(F)}(F)$ for the feature update using its registered mask observations.

The SFS procedure constructs this topology by clustering the lifted feature vectors and encoding the resulting labels as one-hot signals (Xiong et al., 2026). Rendering these signals into a training view and taking the largest response at each pixel gives a projected cluster map. Each SAM mask is assigned to the most frequent cluster label within its region. The intersection over union between the SAM mask and that cluster's full projected region determines whether the registration is retained. The accepted observations are collected across views for each cluster.

Refinement averages the descriptors of the accepted masks within each cluster, assigns that mean to the cluster's Gaussians, and normalizes the updated vectors. Unclassified Gaussians and clusters without accepted masks retain their input features. A mask observation can therefore update Gaussians visible in other views through their common cluster membership.

ToFuGS builds two such fields and uses their training-view support to determine the final fusion weights.

3.2 MASK-INDUCED TOPOLOGY BRANCHES

We form a cross-scale consensus input F_u from all mask granularities and an intermediate-scale input F_m from the predefined hierarchy channel:

$$F_u = \mathrm{norm}\left(\sum_{q \in \mathcal{Q}} \mathrm{norm}(L_q)\right), \quad F_m = \mathrm{norm}(L_{q_{\mathrm{mid}}}). \tag{1}$$

We fix $q_{\mathrm{mid}} = 2$ as a scene-independent intermediate granularity. This branch provides a more localized topology that complements the cross-scale consensus without introducing an additional hierarchy of branches.

The two inputs induce $\mathcal{T}_u = \mathcal{T}(F_u)$ and $\mathcal{T}_m = \mathcal{T}(F_m)$. ToFuGS applies clustering, label projection, mask registration, and refinement once per branch:

$$\widetilde{F}_u = \mathcal{R}_{\mathcal{T}_u}(F_u), \qquad \widetilde{F}_m = \mathcal{R}_{\mathcal{T}_m}(F_m). \tag{2}$$

Each branch retains its own cluster labels and mask registrations. We use *Uniform* and *Mid* for the refined fields $\widetilde{F}_u$ and $\widetilde{F}_m$, respectively; F_u and F_m denote their unrefined inputs. Figure 2 visualizes these branch-specific partitions and compares them with the grouping induced by early fusion.

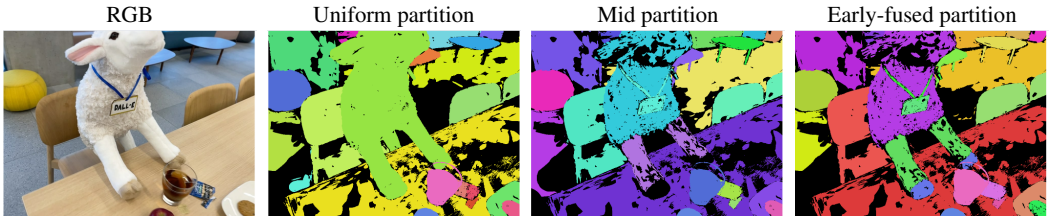


Figure 2: Gaussian partitions on a Teatime view. Uniform and Mid group the same scene differently; mixing their inputs before clustering yields an early-fused partition. Colors denote cluster identities within each panel, not semantic classes.

3.3 Topology before fusion

Let $\mathcal{B}_\alpha(A, B) = \mathrm{norm}((1-\alpha)\,\mathrm{norm}(A) + \alpha\,\mathrm{norm}(B))$ denote row-wise feature fusion. Fusing before refinement yields

$$F_{\text{early}} = \mathcal{R}_{\mathcal{T}(\mathcal{B}_\alpha(F_u, F_m))}(\mathcal{B}_\alpha(F_u, F_m))\,, \tag{3}$$

whereas topology-specific refinement yields

$$F_{\text{late}} = \mathcal{B}_\alpha(\mathcal{R}_{\mathcal{T}_u}(F_u), \mathcal{R}_{\mathcal{T}_m}(F_m))\,. \tag{4}$$

Early fusion removes the identity of the individual branches before their refinement topologies are constructed. As a result, different branch pairs can produce the same fused field even when their separately refined outputs differ. Refining each branch first preserves the information needed to construct its own topology.

3.4 Training-view evidence

For each branch, we ask two questions: do masks assigned to the same cluster remain semantically consistent across training views, and is that cluster supported in both subsets of the views? We measure these two properties as predictive fidelity and registration persistence, respectively. To estimate them, we use J balanced two-way partitions of the training views. For split j, masks registered above the IoU threshold τ contribute trust-weighted descriptors to normalized cluster prototypes $\mu^a_{h,j,c}$ and $\mu^b_{h,j,c}$ in halves a and b. A descriptor x_o from half b, registered to cluster c, is scored against the prototypes from half a:

$$\ell^{a\to b}_{h,j}(o) = \log \frac{\max\Big(1 - \max_{c'\neq c}\cos(x_o, \mu^a_{h,j,c'}), \delta\Big)}{\max\Big(1 - \cos(x_o, \mu^a_{h,j,c}), \delta\Big)}, \qquad \delta > 0. \tag{5}$$

Here δ is a positive distance floor. A positive score indicates that the registered prototype is closer than the nearest competing cluster. The registration-trust-weighted average over scored observations from both directions gives $\bar{\ell}_{h,j,c}$.

Predictive fidelity alone does not indicate whether a cluster is supported across both view halves. We therefore also measure registration persistence. Let $S^a_{h,j,c}$ and $S^b_{h,j,c}$ denote the total trusted registration weights in the two halves:

$$s_{h,j,c} = \sqrt{S^a_{h,j,c} S^b_{h,j,c}}, \qquad e_{h,j,c} = s_{h,j,c}\,[\bar{\ell}_{h,j,c}]_+. \tag{6}$$

Here $[z]_+ = \max(z, 0)$. The geometric mean requires registration support in both halves, while the positive contrast measures how well the cluster predicts descriptors across views. Each Gaussian inherits its cluster's evidence through the branch labels:

$$\rho^{(j)}_{g,h} = 1 + \sqrt{e_{h,j,c_h(g)}}. \tag{7}$$

The square root tempers differences in evidence magnitude before the branch-support ratio is formed. Unsupported or unclassified Gaussians receive unit support. Appendix A.1 specifies the trust weights and prototype construction; Figure 13 shows how the support drives fusion.

3.5 ADAPTIVE FUSION

We convert each split's support scores into a Mid-branch weight $w_{g,j}$. Across the J splits, we compute its mean μ_g, population standard deviation σ_g, and the fraction q_g of splits with evidence for at least one topology:

$$w_{g,j} = \frac{\rho_{g,m}^{(j)}}{\rho_{g,u}^{(j)} + \rho_{g,m}^{(j)}}, \quad \mu_g = \frac{1}{J}\sum_j w_{g,j}, \\ \sigma_g = \mathrm{std}_j(w_{g,j}), \qquad q_g = \frac{1}{J}\sum_j \mathbb{1}\left[\rho_{g,u}^{(j)} > 1 \ \vee \ \rho_{g,m}^{(j)} > 1\right]. \tag{8}$$

The mean μ_g supplies a continuous fusion weight. The fallback selector a_g chooses the branch with greater mean unnormalized support, resolving exact ties by global held-out fidelity (Equation 12 in Appendix A). Here $a_g = 1$ selects Mid and $a_g = 0$ selects Uniform.

The reliability score assesses the continuous weight using evidence coverage, its margin from equal blending, and cross-split variation:

$$m_g = |\mu_g - 0.5|, \qquad r_g = q_g \frac{m_g}{m_g + \lambda_\sigma \sigma_g + \epsilon_r}. \tag{9}$$

A high r_g requires broad evidence coverage and a branch preference that is stable across splits. If $r_g \geq \eta$, we use the continuous weight μ_g. Otherwise, we discard the uncertain mixture weight and select the better-supported branch a_g:

$$\widehat{w}_g = \begin{cases} \mu_g, & r_g \geq \eta, \\ a_g, & r_g < \eta. \end{cases} \tag{10}$$

Inner normalization uses only branch directions and leaves unit-norm features unchanged; outer normalization rescales the weighted sum:

$$F_g^{\mathrm{ToFuGS}} = \mathrm{norm}\Big((1 - \widehat{w}_g)\,\mathrm{norm}(\widetilde{F}_{u,g}) + \widehat{w}_g\,\mathrm{norm}(\widetilde{F}_{m,g})\Big). \tag{11}$$

Both branch fields are refined once; the repeated view partitions estimate their fusion weights. The gate has no learned parameters. Deployment stores $F^{\mathrm{ToFuGS}} \in \mathbb{R}^{N\times 512}$. At an evaluation view, Gaussian rasterization blends these descriptors into a pixelwise feature map. A text query is encoded in the same OpenCLIP space and compared with the rendered features by cosine relevance. The resulting attention map is converted into a binary mask using the readout in Section 4.1. Appendix A lists the construction constants and evaluation steps.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate open-vocabulary segmentation on four LERF scenes (Kerr et al., 2023) and five scenes from the 3D-OVS benchmark introduced by Liu et al. (2023). We report rendered-mask mIoU and average scene scores for the overall mean. The native attention-map protocol converts each query map to a mask using the SFS histogram peak–valley rule: the annotated instance count selects the peak, and an adjacent valley supplies the threshold. Appendix A gives post-processing and the averaging hierarchy.

Main comparisons use three independently trained 3DGS reconstructions, reporting means and, where applicable, SD and paired gains across reconstructions. Unless otherwise stated, ablations use a single geometry, termed the *fixed-reconstruction setting*. The 3D-OVS experiments use the LERF configuration.

ToFuGS uses SFS lifting, refinement, and evaluation (Xiong et al., 2026) with 60k-step 3DGS geometry, SAM ViT-H masks, OpenCLIP ViT-B/16 descriptors, and 512-D fields. The order controls combine F_u and F_m with equal weights, either before refinement (*Fuse→Refine*) or after independent refinement (*Refine→Fuse*). We repeat the same order control with Dr. Splat to test whether the ordering effect depends on the SFS refinement procedure. Topology-sharing variants assign the Uniform or Mid partition to both branches before refinement and adaptive fusion. ToFuGS uses the weights in Equation 11; Appendix A gives implementation details.

4.2 SEGMENTATION PERFORMANCE

Across three independently trained reconstructions, ToFuGS raises the four-scene SFS mean from 61.22 to 68.56 (+7.34; Table 1). It also exceeds both standalone branches, improving on Mid (65.50) by 3.06 points.

Table 1: LERF mIoU (%) across three independently trained reconstructions. Scene entries are reconstruction means; Mean ± SD summarizes the four-scene average across reconstructions. Paired Δ averages within-reconstruction gains over SFS, in percentage points.

Method	Figurines	Ramen	Teatime	Waldo	Mean ± SD	Paired Δ
SFS	64.02	62.48	64.97	53.40	61.22 ± 0.94	–
Uniform	63.71	62.54	71.25	55.46	63.24 ± 2.53	+2.02
Mid	68.14	62.25	**72.88**	58.72	65.50 ± 1.08	+4.28
Refine→Fuse	68.46	64.52	71.53	60.33	66.21 ± 1.96	+5.00
ToFuGS	**71.07**	**66.22**	72.34	**64.62**	$\mathbf{68.56 \pm 1.86}$	**+7.34**

The largest scene gain over SFS occurs on Waldo Kitchen. In Figure 3, ToFuGS recovers the plastic ladle and more of the refrigerator region than SFS.

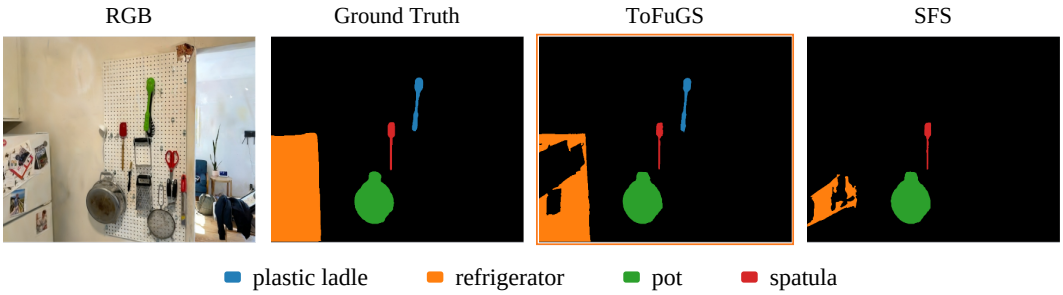


Figure 3: Qualitative segmentation on a Waldo Kitchen view in LERF. ToFuGS recovers the plastic ladle and more of the refrigerator region than SFS. Colors denote the semantic classes in the legend.

Across three independently trained reconstructions, ToFuGS reaches 68.56 mIoU, ahead of LEGO at 67.18 (Table 2a). The two methods have different scene profiles: LEGO scores highest on Ramen and Teatime, while ToFuGS leads on Figurines and Waldo Kitchen. Appendix G also compares scores reported in the original papers (Table 9).

On 3D-OVS, the LERF configuration yields 74.03 ± 0.91 mIoU for ToFuGS, compared with 69.69 ± 0.70 for SFS (Table 2c). The gains are 4.34 ± 1.26 points over SFS and 1.00 ± 0.29 over Refine→Fuse.

4.3 OPERATOR ORDER AND BRANCH-SPECIFIC TOPOLOGY

Changing operator order from Fuse→Refine to Refine→Fuse raises the LERF mean from 64.19 to 66.64 (+2.45; Table 2d). The same ordering improves all four scenes with a Dr. Splat backend (Kim et al., 2025), raising its mean by 2.78 points (Table 2d; Appendix F).

Figure 4 shows the changes in two query masks. For "spoon," Fuse→Refine labels part of the surrounding bowl as spoon, while Refine→Fuse confines the foreground prediction more closely to the utensil. For "sink," the main change is in coverage: Refine→Fuse recovers basin regions missed by Fuse→Refine, although it also includes the cups inside the basin.

The independent-topology configuration exceeds the shared-topology variants by 1.93–3.54 mean mIoU points (Table 2b), leading in three scenes; Shared-$\mathcal{T}_u$ scores highest on Ramen.

Table 2: **Segmentation results and ablations.** Scores are mIoU (%); Mean denotes the average across scenes. Panels (a, c) summarize three independently trained reconstructions; (b, d) use the fixed-reconstruction setting. F, R, T, and W denote Figurines, Ramen, Teatime, and Waldo Kitchen. Bold indicates the best result in each comparison.

(a) Comparison with existing methods on LERF. We compare six methods on four LERF scenes using the attention-map readout in Section 4.1. Scores are reconstruction means; SFS and ToFuGS use the values from Table 1.

Method	F	R	T	W	Mean
SFS	64.02	62.48	64.97	53.40	61.22
OpenGaussian	56.27	39.37	66.60	47.15	52.35
CCL-LGS	38.78	53.79	52.88	57.00	50.61
LangSplatV2	44.62	45.27	52.42	43.42	46.43
LEGO	69.25	**67.27**	**78.68**	53.53	67.18
ToFuGS	**71.07**	66.22	72.34	**64.62**	**68.56**

(b) Topology sharing on LERF. Independent (ToFuGS) uses separate refinement topologies for the Uniform and Mid branches, whereas Shared-$\mathcal{T}_u$ and Shared-$\mathcal{T}_m$ reuse the Uniform or Mid topology across both branches. Both branches are refined before adaptive fusion.

Topology	F	R	T	W	Mean
Independent	**71.77**	64.53	**74.44**	**61.80**	**68.14**
Shared-$\mathcal{T}_u$	61.24	**66.10**	72.13	58.93	64.60
Shared-$\mathcal{T}_m$	66.12	63.64	74.14	60.91	66.20

(c) Transfer to 3D-OVS. Mean $\pm$ SD is computed over three reconstructions; paired Δ denotes the gain over SFS in percentage points.

Method	Mean $\pm$ SD	Paired Δ
SFS	69.69 ± 0.70	–
Mid	72.63 ± 0.58	$+2.94$
Refine→Fuse	73.04 ± 0.62	$+3.35$
ToFuGS	$\mathbf{74.03 \pm 0.91}$	$\mathbf{+4.34}$

(d) Order and fusion on LERF. Operator-order controls use equal-weight fusion with both backends; the SFS variants compare fusion rules and the number of view splits.

Backend	Variant	Mean
SFS	Fuse→Refine	64.19
	Refine→Fuse	66.64
	Multi-split mean	66.79
	Stability-weighted soft	67.21
	Multi-split MAP	67.28
	ToFuGS, one split	67.35
	ToFuGS, five splits	**68.14**
Dr. Splat	Fuse→Refine	49.56
	Refine→Fuse	**52.34**

Sharing partitions also makes the branch evidence identical, so the gate selects Mid throughout all four scenes. Independent topologies retain soft blending for 7.9%–13.8% of Gaussians. This ablation changes both the refinement topology and the fusion behavior it induces.

The shared variants differ because Mid is refined under either its own partition or Uniform's partition. Appendix F gives the construction details.

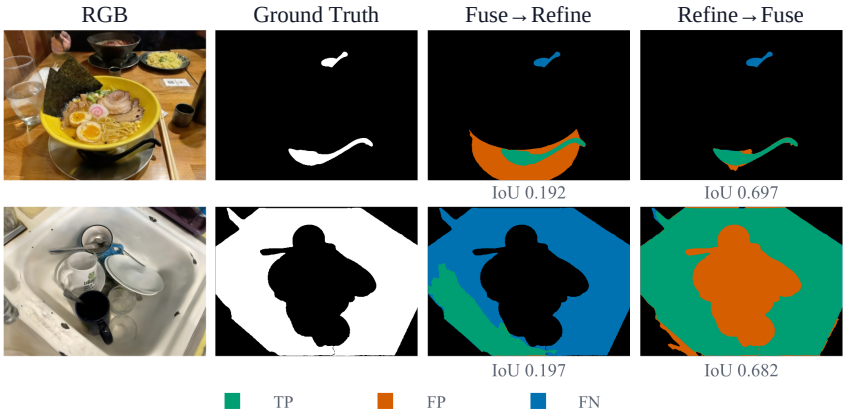


Figure 4: Operator order on Ramen "spoon" (top) and Waldo "sink" (bottom). Refine→Fuse reduces false positives and recovers missed regions. Green, orange, and blue mark true positives, false positives, and false negatives; scores are query-level IoU on $[0, 1]$.

4.4 Adaptive fusion

Adaptive fusion improves over equal-weight Refine→Fuse by 1.50 mIoU points, reaching 68.14 (Table 2d). Simply averaging the split-wise weights provides little benefit: when different splits prefer different branches, their mean produces an intermediate mixture. Stability-aware alternatives perform better by moving toward a single branch when the preference is inconsistent across splits. ToFuGS retains continuous weights in reliable regions and selects a branch elsewhere.

Uniform and Mid make different local errors (Figure 5). For "sink," Mid separates the basin from the cup and plate that Uniform includes. For "sheep," Uniform covers the legs that Mid largely omits. ToFuGS preserves both the exclusion of objects from the sink region and the coverage of the sheep's legs, approaching the better branch's IoU in each example.



Figure 5: Complementary branch predictions. ToFuGS retains Mid's coverage for "sink" and Uniform's for "sheep." Orange frames mark the higher-IoU branch; pink overlays and green contours indicate predictions and ground truth. Scores are query-level IoU on $[0, 1]$.

Removing cross-split variation or the weight margin lowers segmentation mIoU by 0.42 or 0.45 points, respectively. Appendix D reports these variants and the relation between reliability and branch support on omitted view splits. Additional parameter and boundary analyses appear in Appendices B and E.

4.5 Deployment cost

Despite its topology-specific refinement and evidence estimation, ToFuGS remains lightweight at deployment, retaining the SFS footprint of one 970.73 MiB field and one render. Ramen evaluation takes 43.99 s with 2.36 GiB peak host memory, compared with 49.98 s and 2.35 GiB for SFS. On a single RTX 3090, construction from cached lifted features takes 23 min 45 s for Ramen: about 5 min 7 s for the two branch refinements and 18 min 38 s for evidence estimation and fusion. The five view partitions account for ten topology–split evidence passes. Appendix A.2 gives the timing breakdown. Replacing Mid with hierarchy-wide routing adds 1.13 points on a separate reconstruction, at approximately 4.00× the semantic storage and 5.00× the runtime (Table 5).

5 Limitations and Scope

ToFuGS relies on projected clusters that overlap SAM masks across views and on an intermediate-scale mask channel that separates object parts or compact instances.

A nonzero fusion weight can carry errors from Uniform into a region where Mid is more accurate. On Teatime, Mid retains a scene mean of 72.88 mIoU compared with 72.34 for ToFuGS. The training-

view evidence measures agreement among registered descriptors, so consistent mask or descriptor errors can also receive strong support.

6 Conclusion

ToFuGS shows that fusion order matters in open-vocabulary Gaussian splatting. It improves SFS by 7.34 mIoU points on LERF, transfers the same configuration to 3D-OVS, and retains a single-descriptor deployment. These results suggest that multiscale semantic fields should preserve their feature-induced refinement structures until refinement is complete, rather than being fused beforehand.

Reproducibility Statement

The anonymous repository at `https://anonymous.4open.science/r/ToFuGS-E146` provides the core implementation of independent refinement, training-view evidence estimation, and adaptive fusion. All model weights and the complete source code will be released upon acceptance. Initial 3DGS checkpoints can be constructed using the standard procedure described in Appendix A, which also documents the evaluation protocol. Appendix A.2 reports resource and rebuild checks.

The experiments use the public LERF and 3D-OVS scene collections. Derived masks, lifted features, predictions, and evaluation summaries will be distributed subject to the source datasets' licenses. No private evaluation data are used.

AI Use Statement

Generative AI tools were used to assist with literature organization, language polishing, and LaTeX formatting. The authors are responsible for verifying all technical claims, citations, experimental results, and the final manuscript.

References

Jiazhong Cen, Xudong Zhou, Jiemin Fang, Changsong Wen, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Tackling view-dependent semantics in 3D language gaussian splatting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7013–7034, 2025. URL `https://proceedings.mlr.press/v267/cen25a.html`.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023.

Jun-Seong Kim, GeonU Kim, Yu-Ji Kim, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. Splat: Directly referring 3D gaussian splatting via direct language embedding registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14137–14146, 2025. doi: 10.1109/CVPR52734.2025.01319.

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

Wanhua Li, Yujie Zhao, Minghan Qin, Yang Liu, Yuanhao Cai, Chuang Gan, and Hanspeter Pfister. LangSplatV2: High-dimensional 3D language gaussian splatting with 450+ FPS. In *Advances in Neural Information Processing Systems*, 2025. URL `https://arxiv.org/abs/2507.07136`.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 53433–53456, 2023. doi: 10.52202/075280-2325.

Ziqi Lu, Heng Yang, Danfei Xu, Boyi Li, Boris Ivanovic, Marco Pavone, and Yue Wang. LoRA3D: Low-rank self-calibration of 3D geometric foundation models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LSp4KBhAom.

Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3D vision-language gaussian splatting. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SSE9myD9SG.

Yuning Peng, Haiping Wang, Yuan Liu, Yipeng Lu, Zhen Dong, and Bisheng Yang. LEGO: Leveled language gaussian splatting. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2608.10057.

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 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3D gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5333–5343, 2024.

Lei Tian, Xiaomin Li, Liqian Ma, Hao Yin, Zirui Zheng, Hefei Huang, Taiqing Li, Huchuan Lu, and Xu Jia. CCL-LGS: Contrastive codebook learning for 3D language gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2505.20469.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. OpenGaussian: Towards point-level 3D gaussian-based open vocabulary understanding. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.02058.

Butian Xiong, Rong Liu, Kenneth Xu, Meida Chen, and Andrew Feng. Splat feature solver. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AepuXqQM4X.

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.

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.

A Implementation and Protocol Details

Internal variants share 60k-step 3DGS geometry, SAM ViT-H masks, OpenCLIP ViT-B/16 descriptors, and the published SFS clustering, refinement, and evaluation procedures. Field construction uses only training RGB images, masks, and descriptors. Text queries and held-out annotations are used for evaluation. The same predefined intermediate-scale channel supplies F_m across scenes, reconstructions, and datasets, and its refined field $\widetilde{F}_m$ is denoted by *Mid*.

Attention-map evaluation. The native attention-map protocol applies the SFS histogram peak–valley rule to each query map. The annotated instance count for the query determines which histogram peak is selected, and an adjacent valley determines the threshold. When the peak–valley rule yields no threshold, the evaluator thresholds a nearly constant map at its mean and otherwise averages the Otsu threshold with the 75th percentile. Binary masks undergo 3×3 morphological smoothing, largest-component retention, and hole filling. IoU is averaged over query categories within each view, over labeled views within each scene, and then equally over scenes. Reported gains are computed from unrounded scores.

Table 3a lists the defaults. The shared-topology experiment uses reconstruction seed 50. The five view partitions consist of one sorted-view parity split and four deterministically hashed splits with fixed keys. Applied to both topologies, they yield ten topology–split evidence passes. The local sweeps in Appendix B vary one reliability parameter at a time while retaining this default configuration for the benchmark evaluations.

Table 3: Construction defaults and deployment cost. Defaults apply across all benchmarks. Ramen timings cover evaluation of the deployed field with lifting reused; RSS is peak host memory.

(a) Construction constants

Quantity	Symbol	Value
View partitions	J	5
Registration IoU threshold	τ	0.8
Contrastive distance floor	δ	10^{-4}
Variation coefficient	λ_σ	2
Reliability offset	ϵ_r	0.02
Soft/hard threshold	η	0.5
Evidence pseudocount	–	1
Tikhonov coefficient	–	1.2

(b) Ramen evaluation

Method	Time (s)	RSS (GiB)	Field (MiB)
SFS	49.98	2.35	970.73
Mid	42.47	2.36	970.73
ToFuGS	43.99	2.36	970.73

Fallback branch selection. Let $\bar{\rho}_{g,h} = J^{-1} \sum_j \rho^{(j)}_{g,h}$ be the mean support for Gaussian g under topology h, and let Φ_h be topology h's split-averaged global held-out log score, defined in Appendix A.1. The fallback selects the branch with greater mean support, using Φ_h to break ties:

$$a_g = \begin{cases} 1, & \bar{\rho}_{g,m} > \bar{\rho}_{g,u}, \\ 0, & \bar{\rho}_{g,m} < \bar{\rho}_{g,u}, \\ \mathbf{1}[\Phi_m \geq \Phi_u], & \bar{\rho}_{g,m} = \bar{\rho}_{g,u}. \end{cases} \tag{12}$$

A.1 Training-view evidence estimator

For each topology $h \in \{u, m\}$, we project the partition c_h into each training view and register mask observation o, with descriptor x_o, to cluster $c(o)$ at IoU t_o. Observations with $t_o > \tau$ receive the trust weight

$$\omega_o = \frac{t_o - \tau}{1 - \tau}. \tag{13}$$

For cluster c in side a of split j, the normalized prototype is

$$\mu^a_{h,j,c} = \operatorname{norm}\left(\sum_{o \in \mathcal{O}^a_{h,j,c}} \omega_o \operatorname{norm}(x_o) \right), \tag{14}$$

where $\mathcal{O}^a_{h,j,c}$ contains the trusted observations assigned to c on side a; side b is analogous. A prototype is valid only if its total trust weight is positive and the norm of its weighted descriptor sum is finite and greater than 10^{-12}. Invalid prototypes are omitted. Equation 5 scores an observation only when the opposite half has a valid prototype for its assigned cluster and at least one other valid prototype; the nearest competitor is selected from these valid alternatives. Otherwise, the observation is skipped.

Let $\mathcal{E}_{h,j,c}$ contain the scored observations assigned to c from both directions, and let $\ell_{h,j}(o)$ denote each observation's opposite-half score. The pooled contrast is a weighted average:

$$W_{h,j,c} = \sum_{o \in \mathcal{E}_{h,j,c}} \omega_o, \qquad \bar{\ell}_{h,j,c} = \frac{\sum_{o \in \mathcal{E}_{h,j,c}} \omega_o \ell_{h,j}(o)}{W_{h,j,c}} \quad (W_{h,j,c} > 0).$$

When $W_{h,j,c} = 0$, the cluster contributes zero evidence and its Gaussians receive unit support. The trusted registration mass is $S^a_{h,j,c} = \sum_{o \in \mathcal{O}^a_{h,j,c}} \omega_o$, with $S^b_{h,j,c}$ defined analogously. Zero mass in either half also gives zero persistence and evidence.

For the global tie-breaking score, let $\mathcal{C}_{h,j} = \{c : W_{h,j,c} > 0\}$ be the scored clusters. We average their signed pooled contrasts equally within each split, then average across splits:

$$\Phi_h = \frac{1}{J} \sum_{j=1}^{J} \frac{\sum_{c \in \mathcal{C}_{h,j}} \bar{\ell}_{h,j,c}}{\max(1, |\mathcal{C}_{h,j}|)}.$$

An empty scored-cluster set contributes zero to this average.

A.2 Efficiency and reproducibility

ToFuGS retains the same 970.73 MiB deployed field size as SFS, with comparable evaluation memory (Table 3b). Offline construction peaks at approximately 9.98 GiB allocated and 11.57 GiB reserved GPU memory. Starting from cached lifted fields on Ramen, Uniform refinement takes 156.52 s and Mid refinement takes 150.95 s on one RTX 3090. Evidence estimation and final fusion take 1117.77 s, giving 1425.24 s from lifted inputs to the exported field. This interval covers branch refinement and fusion; geometry reconstruction and initial lifting precede it.

Three independent rebuilds produce identical SHA-256 hashes for the refined Uniform, refined Mid, and final ToFuGS tensors, with zero maximum absolute difference. The anonymous repository provides the core refinement and fusion implementation.

B Reliability Hyperparameter Sensitivity

We vary one parameter around $(J, \eta, \tau, \lambda_\sigma) = (5, 0.5, 0.8, 2.0)$ on the fixed reconstruction, keeping geometry, refined branch fields, labels, queries, and the official SFS attention-map evaluator unchanged. The default four-scene mean is 68.1373 mIoU.

Performance nearly saturates by $J = 3$ view splits (Figure 6). The four-scene mIoU spans 0.15 points for $\eta \in [0.4, 0.6]$ and 0.13 points for $\lambda_\sigma \in [1, 3]$. The registration threshold has a larger effect: changing τ from 0.8 to 0.7 or 0.9 lowers mIoU by 0.97 or 0.85 points, respectively, and can switch a scene's fallback branch. This difference follows the roles of the parameters: η and λ_σ adjust the use of an estimated weight, whereas τ changes the registered observations from which the evidence is built.

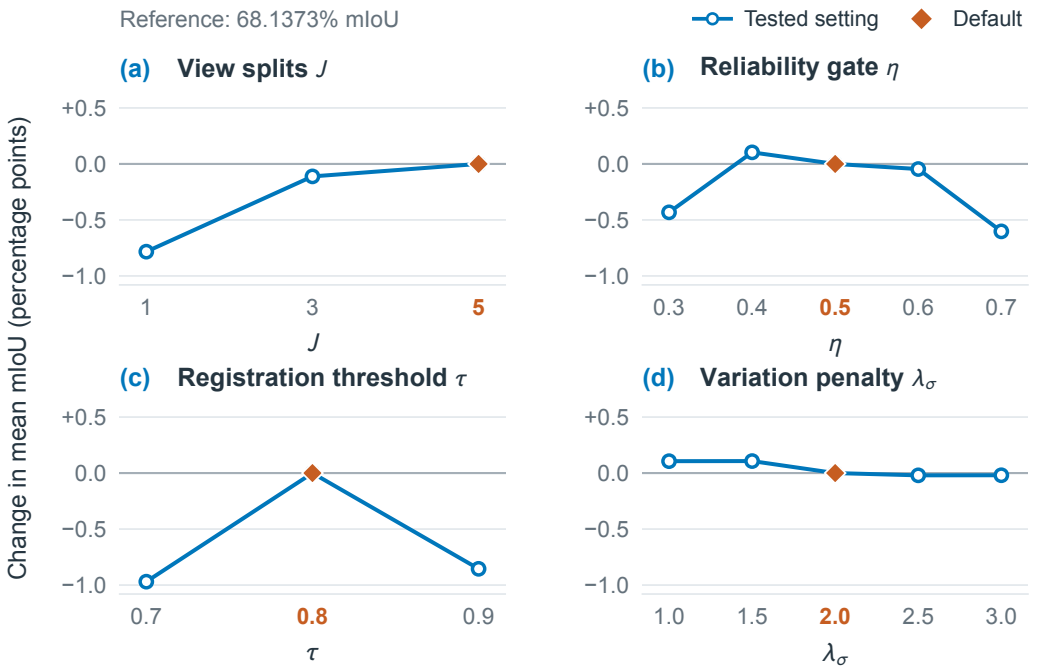


Figure 6: Local reliability sensitivity on the fixed LERF reconstruction. Lines connect tested settings and show changes in four-scene mean mIoU (percentage points) relative to the default configuration (68.1373%). Orange diamonds mark the default value in each one-variable sweep; all panels share the same vertical scale.

C Complementarity Diagnostics

Uniform and Mid induce different boundaries and region assignments, while early fusion constructs a third partition from their mixed features (Figures 2 and 7). Table 4 quantifies this structural complementarity across three independently trained 3DGS reconstructions. Mid reaches 65.50 mIoU; its partition has normalized mutual information (NMI) 0.840 with Uniform's partition. ToFuGS combines these branches to reach 68.56 and exceeds the Equal, MAP, and Hard fusion controls.

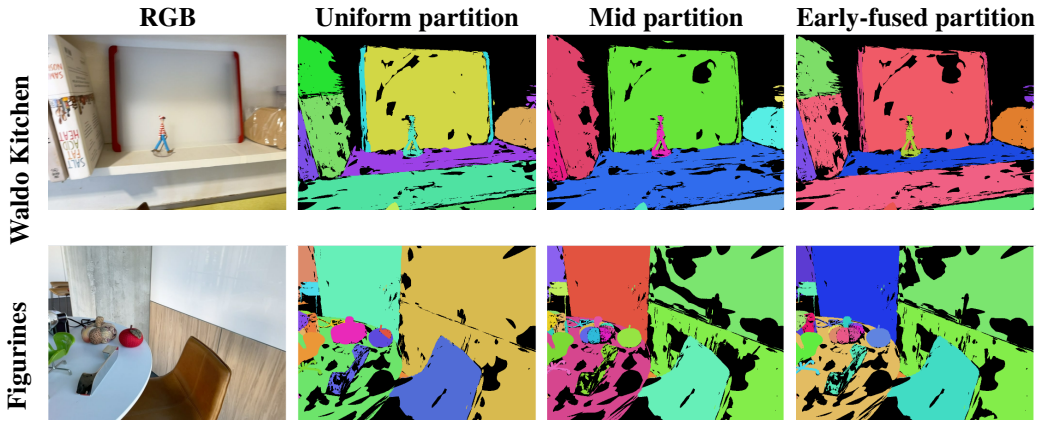


Figure 7: Additional Gaussian partitions induced by Uniform, Mid, and early-fused features in Waldo Kitchen and Figurines. Colors denote partition identities within each panel, not semantic classes. The panels expose differences in region boundaries and assignments before fusion.

Table 4: Mid complementarity on LERF across three independent reconstructions. NMI compares the Mid and Uniform partitions; the fusion controls combine the two refined fields.

	Branch diagnostics		Fusion mIoU (%)		
Branch	mIoU (%)	NMI	Equal	MAP	Hard
Mid	65.50	0.840	66.21	65.87	65.93

Hierarchy route pairs Uniform with a second branch that routes across the full mask hierarchy in place of Mid. On a separate reconstruction (Table 5), it raises the mean from 68.98 to 70.11, with gains concentrated in Teatime and Waldo Kitchen. This 1.13-point gain requires approximately $4.00\times$ the semantic-field storage and $5.00\times$ the runtime. The Mid branch thus offers a more compact operating point.

Table 5: Hierarchy routing on one shared LERF reconstruction, distinct from the fixed-reconstruction analyses. Scores are mIoU (%), and Δ is Hierarchy route minus ToFuGS. Storage and runtime are relative to ToFuGS.

Variant	Figurines	Ramen	Teatime	Waldo	Mean	Storage	Runtime
ToFuGS	**73.09**	**65.05**	74.91	62.87	68.98	$1.00\times$	$1.00\times$
Hierarchy route	72.73	64.54	**77.25**	**65.90**	**70.11**	$\sim 4.00\times$	$\sim 5.00\times$
Δ	-0.36	-0.50	$+2.34$	$+3.03$	$+\mathbf{1.13}$	–	–

D Reliability and Fusion Variants

Fusion rules. With the independently refined Uniform and Mid fields held fixed, five-split ToFuGS reaches 68.14 mIoU (Table 2d). It exceeds equal-weight late fusion by 1.50 points and its one-split counterpart by 0.79 points, separating the effects of the fusion rule and repeated evidence estimation.

Removing cross-split variation or the weight margin lowers the mean by 0.42 or 0.45 points, respectively. These terms distinguish a consistent branch preference from a large average weight that changes with the view partition.

The controls differ in the Mid weight used in Equation 11. Equal-weight Refine→Fuse uses $\widehat{w}_g = 0.5$; Multi-split mean uses μ_g; Stability-weighted soft uses $a_g + r_g(\mu_g - a_g)$; and Multi-split MAP uses the fallback branch a_g. ToFuGS applies the thresholded rule in Equation 10, with $J = 1$ for its one-split variant and $J = 5$ for the full construction.

Reliability diagnostics. For each omitted split j, we recompute reliability $r_g^{(-j)}$ and select a candidate topology $h_g^{(-j)}$ using mean supports over the other $J - 1$ splits and the fixed global tie-breaking rule. We then compare the two support scores from split j. Writing $\bar{h}_g^{(-j)}$ for the other topology, the normalized precision advantage is

$$A_{g,j} = \frac{\rho^{(j)}_{g,h_g^{(-j)}} - \rho^{(j)}_{g,\bar{h}_g^{(-j)}}}{\rho^{(j)}_{g,h_g^{(-j)}} + \rho^{(j)}_{g,\bar{h}_g^{(-j)}}}. \tag{15}$$

Within bins of $r_g^{(-j)}$, we report the fraction with $A_{g,j} > 0$ and the mean of $A_{g,j}$, pooling Gaussian–split pairs across all four scenes. The diagnostic measures whether the candidate topology's preference carries over to an omitted evidence split; the deployed field uses the soft/hard weights in Equation 10. Table 6 reports the bin summaries, and Figure 8 plots them with bin counts.

Table 6: Reliability and omitted-split support on LERF, pooled over Gaussian–split pairs. Win rate is the fraction with $A_{g,j} > 0$; normalized advantage is defined in Equation 15.

Reliability interval	Support win rate (%)	Mean normalized advantage
$[0, 0.1)$	4.1	0.001
$[0.1, 0.25)$	81.5	0.030
$[0.25, 0.5)$	94.9	0.070
$[0.5, 0.75)$	92.9	0.167
$[0.75, 1.0]$	99.4	0.499

Mean normalized advantage increases from 0.001 in the lowest reliability bin to 0.499 in the highest. The highest-reliability candidates win 99.4% of omitted-split comparisons.

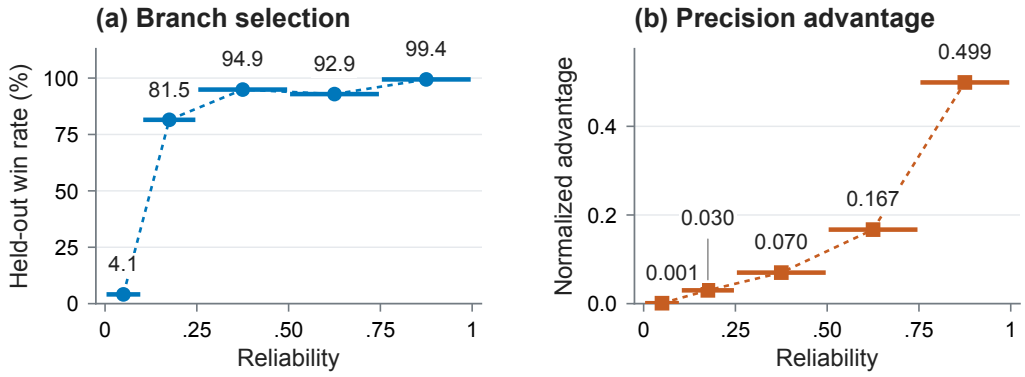


Figure 8: Leave-one-split-out evidence diagnostics pooled across four LERF scenes on the fixed reconstruction: (a) candidate-branch win rate and (b) normalized precision advantage. Horizontal segments span reliability bins; dashed lines join bin summaries. Counts of Gaussian–split pairs, from left to right, are 18.88, 1.16, 1.94, 1.25, and 1.56 million.

E Boundary and Size-Stratified Results

Boundary-IoU uses a tolerance of 2% of the image diagonal. We group targets by ground-truth mask area relative to image area: Small ($< 2\%$), Medium (2%–$< 10\%$), and Large ($\geq 10\%$). ToFuGS improves over Mid in every group on the fixed reconstruction (Figure 9).

On the fixed reconstruction, ToFuGS improves over Mid by 1.94 mIoU and 2.13 Boundary-IoU points. Small targets show the largest mIoU gain ($+2.51$), and large targets the largest boundary gain ($+3.36$). For large targets, the boundary gain exceeds the full-mask IoU gain of 1.56 points, indicating a larger improvement in contour agreement than in overall region overlap. Improvements on medium targets are smaller ($+0.20$ mIoU and $+0.09$ Boundary-IoU). Table 7 reports the absolute scores for each group.

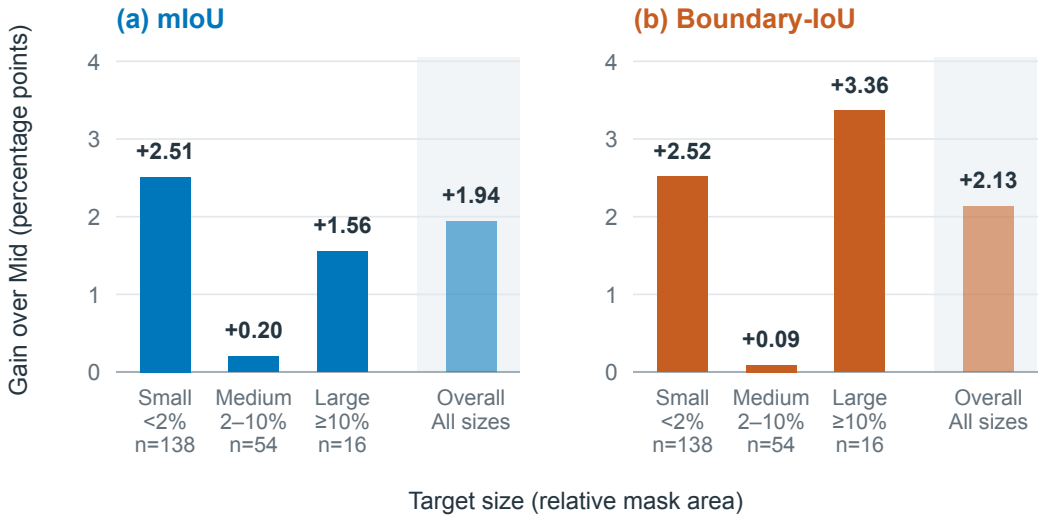


Figure 9: ToFuGS-minus-Mid gains on the fixed LERF reconstruction: (a) mIoU and (b) Boundary-IoU, in percentage points on a shared vertical scale. Targets are grouped by mask area relative to image area; n gives the instance count. Shaded columns separate the overall summary from the three size groups. Gains are computed from unrounded scores.

Table 7: Absolute size-stratified scores on the fixed LERF reconstruction. Both metrics are reported in percent; n is the number of target instances.

		IoU		Boundary-IoU	
Subset	n	Mid	ToFuGS	Mid	ToFuGS
Small	138	61.82	**64.33**	60.65	**63.16**
Medium	54	71.43	**71.63**	62.84	**62.93**
Large	16	79.04	**80.60**	67.08	**70.44**

F Topology and Operator-Order Analysis

Refine→Fuse outperforms Fuse→Refine in all four scenes with both refinement backends (Table 8). The mean gains are 2.45 points for SFS and 2.78 for Dr. Splat.

Dr. Splat adaptation. We retain Dr. Splat's registration, normalization, quantization, and rendering. Uniform and Mid are refined separately before equal-weight fusion, or fused first and then refined.

Table 8: Per-scene LERF mIoU (%) for SFS (top, fixed reconstruction) and Dr. Splat (bottom). Within each backend, the two controls share equal-weight fusion and differ only in operator order.

Method	Figurines	Ramen	Teatime	Waldo	Mean
Primary backend: SFS					
Uniform	61.17	64.13	71.85	55.03	63.04
Mid	66.12	63.64	74.14	60.91	66.20
Fuse→Refine	63.75	65.53	71.08	56.39	64.19
Refine→Fuse	66.75	**65.67**	73.28	60.88	66.64
ToFuGS	**71.77**	64.53	**74.44**	**61.80**	**68.14**
Second backend: Dr. Splat					
Fuse→Refine	39.96	54.80	49.72	53.78	49.56
Refine→Fuse	**42.89**	**57.32**	**55.22**	**53.92**	**52.34**

Shared-topology construction. The control in Table 2b uses the fixed LERF reconstruction and input fields F_u and F_m. Independent refinement uses $(\mathcal{T}_u, \mathcal{T}_m)$. Shared-$\mathcal{T}_u$ reuses the Uniform partition for both branches, and Shared-$\mathcal{T}_m$ reuses the Mid partition. Each branch undergoes mask registration and refinement under its assigned partition, followed by five-split evidence estimation and adaptive fusion.

With shared partitions and registrations, $\rho_{g,u}^{(j)} = \rho_{g,m}^{(j)}$, so $\mu_g = 0.5$, $m_g = 0$, and $r_g = 0$. The gate therefore takes the hard-selection path, whose tie-breaking selects the Mid-input branch in all four scenes. This branch is refined under the shared partition. The independent configuration retains a nonzero soft-blending region in every scene, spanning 7.9%–13.8% of Gaussians at the fixed threshold $\eta = 0.5$.

A real-scene non-commuting example. We compare $\mathcal{T}_u$, $\mathcal{T}_m$, and the early-fusion topology $\mathcal{T}_{\text{mix}}$ in a 256×256 ROI from Waldo Kitchen, frame `00056`. The camera is fixed using a training-only score combining topology-boundary disagreement and registration stability; within that view, the crop maximizes three-way topology-boundary disagreement. All three induced partitions differ. We then compare the normalized descriptors from Fuse→Refine and equal-weight Refine→Fuse on the unique Gaussians that win the visibility assignment at pixels in this crop. Their cosine similarity is below 0.999 for 40.62% of these Gaussians, and their mean angular difference is $5.30°$.

G Additional Evaluation and Qualitative Results

Comparison of original reported scores. Table 9 collects the LERF scores reported in each method's original paper. ToFuGS reaches 68.56 mIoU using the three-reconstruction means from Table 1.

Table 9: LERF mIoU (%) reported in the original papers, using each method's reported protocol. ToFuGS uses the three-reconstruction means from Table 1. Waldo denotes Waldo Kitchen.

Method	Figurines	Ramen	Teatime	Waldo	Mean
LangSplat (Qin et al., 2024)	44.7	51.2	65.1	44.5	51.4
OpenGaussian (Wu et al., 2024)	39.29	31.01	60.44	22.70	38.36
LaGa (Cen et al., 2025)	64.1	55.6	70.9	65.6	64.0
LangSplatV2 (Li et al., 2025)	56.4	51.8	72.2	59.1	59.9
CCL-LGS (Tian et al., 2025)	61.2	62.3	71.8	**67.1**	65.6
SFS (Xiong et al., 2026)	67.6	62.3	68.5	62.1	65.1
LEGO (Peng et al., 2026)	62.3	**67.5**	**81.0**	62.7	68.4
ToFuGS	**71.07**	66.22	72.34	64.62	**68.56**

Additional scene comparisons. Figure 10 compares Ramen, Teatime, and Figurines. ToFuGS preserves more complete target regions than SFS and CCL-LGS in the Ramen and Teatime views, and recovers the pumpkin label missed by SFS in Figurines. The Waldo Kitchen comparison is shown in Figure 3.

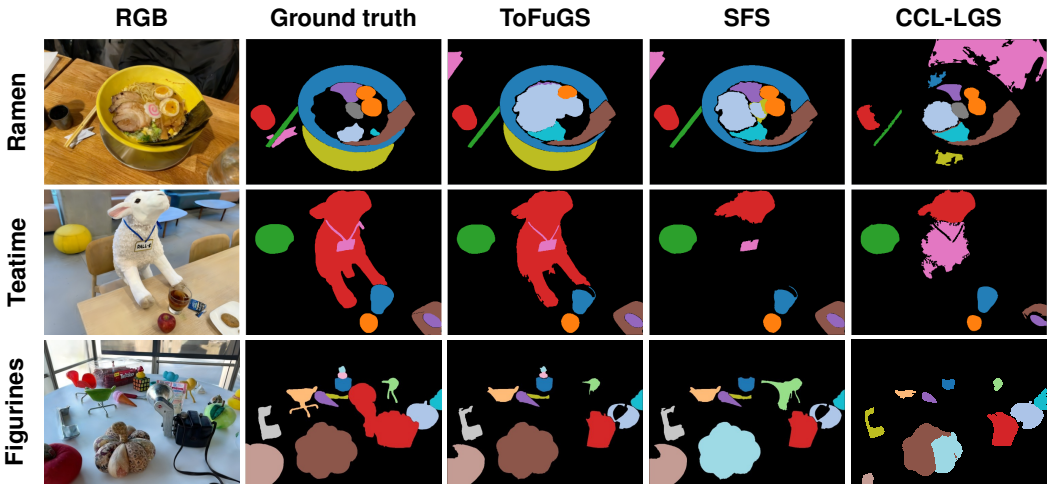


Figure 10: Additional LERF comparisons on Ramen, Teatime, and Figurines. Columns show RGB, ground truth, ToFuGS, SFS, and CCL-LGS. Colors distinguish semantic classes within each scene.

Cross-view behavior. For the Figurines query "rubber duck with hat," ToFuGS recovers the target in three views missed by SFS and remains comparable in the fourth (Figure 11).

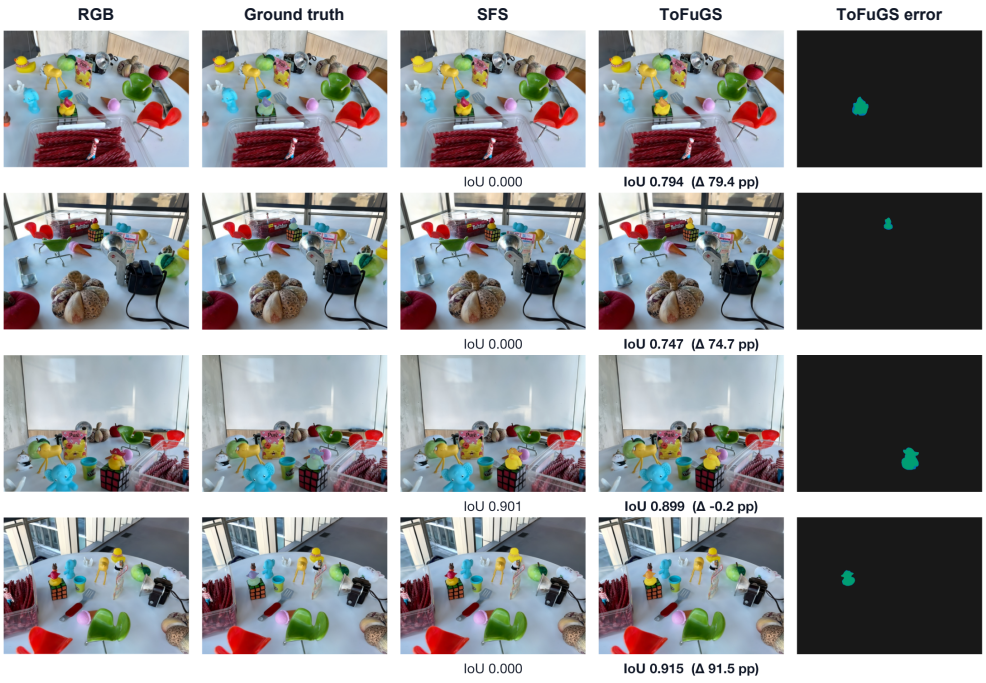


Figure 11: Cross-view IoU for "rubber duck with hat" in Figurines across all four labeled views containing the target; Δ is ToFuGS minus SFS, in percentage points. Green, vermilion, and blue mark true positives, false positives, and false negatives, respectively.

Query-level diagnostics. Figure 12 links ToFuGS's coverage gains for refrigerator and spatula queries to attention responses and IoU under the official SFS attention-map protocol.

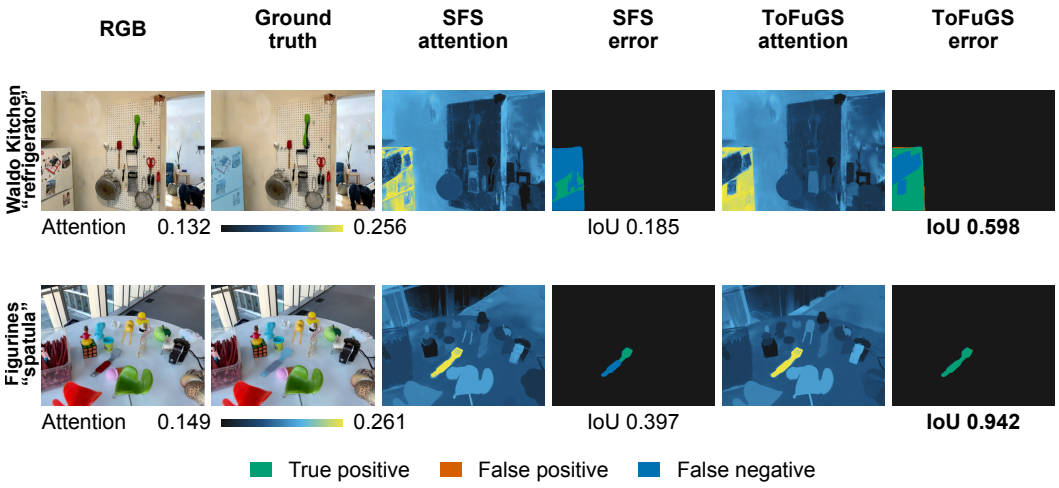


Figure 12: Selected high-gain queries on the fixed LERF reconstruction, evaluated with the official SFS attention-map protocol. Each row shares an attention range set by both methods' joint 1st and 99th percentiles. Error maps mark true positives, false positives, and false negatives; scores are query-level IoU.

H Adaptive Fusion in Detail

Figure 13 expands panel (c) of Figure 1, separating evidence estimation from the feature data path. Both branch fields are refined once; view resampling estimates only their fusion weights.

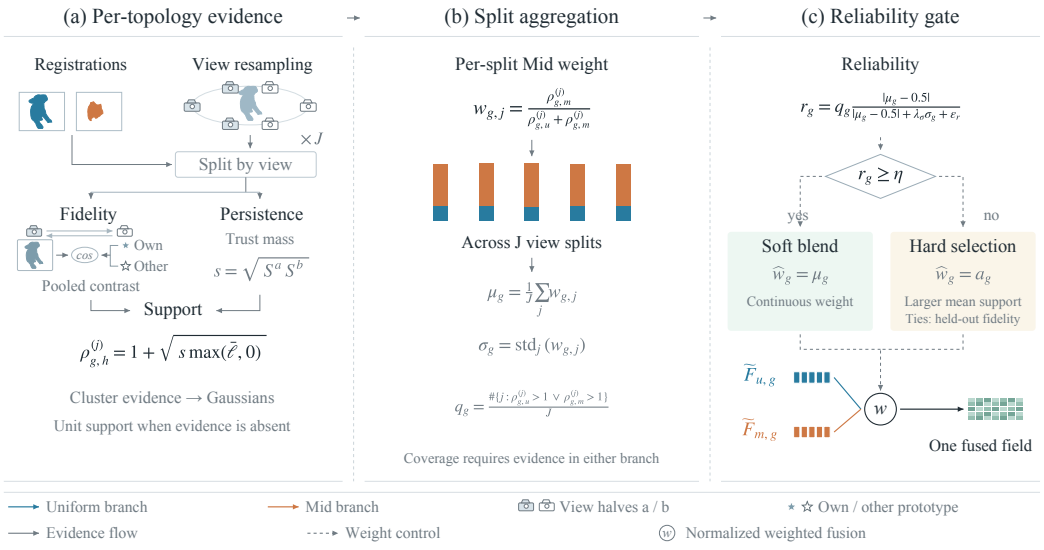


Figure 13: Evidence-guided fusion. (a) Opposite-half prototype contrasts and trusted registration mass produce branch supports. (b) Split-wise weights yield their mean, variation, and evidence coverage. (c) Reliability selects soft blending or a hard branch before normalized fusion. Indices $(h, j, c_h(g))$ are suppressed in s and $\bar{\ell}$; bars are schematic. The hard-branch tie rule follows Equation 12.