GEOMETRY-AWARE FEATURE FUSION FOR OPEN-VOCABULARY 3D SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary 3D segmentation relies on integration of multi-view language features into a spatially coherent scene representation. Although existing language and instance representations provide semantic and object-level information, their contributions to individual 3D elements may become ambiguous, since region-level observations can span neighboring objects or cover only parts of a target. To address this issue, this work presents Geometry-Aware Feature Fusion (GAFF), a framework that uses geometric relationships and instance structure to guide the integration of language evidence into a 3D representation, without requiring additional gradient-based optimization. Geometrically admissible observations are weighted according to depth consistency, local region agreement, and instance compatibility, then combined with the existing scene semantics. This formulation grounds semantic aggregation in both local geometry and object-level structure. Although this fusion refines scene semantics, language relevance alone does not determine the correct spatial extent of a queried object. We therefore further use object-guided mask refinement to select among multilevel language masks and adjust their extent using object candidates. GAFF achieves 68.45% mIoU on LERF-OVS and 94.88% on 3D-OVS. On a ten-scene subset of ScanNetV2, it achieves 38.28% mIoU for 19-class point-level semantic segmentation. Controlled fusion experiments demonstrate improved semantic prediction without mask editing, while component studies show complementary benefits from feature fusion and mask refinement. These results highlight the value of using geometry and instance structure to organize multi-view semantic evidence for open-vocabulary 3D segmentation.

1 INTRODUCTION

Open-vocabulary 3D segmentation associates natural-language queries with spatial regions beyond a predefined training taxonomy. Recent methods integrate language-aligned semantics into 3D Gaussian scene models for flexible scene understanding and interaction (Ding et al., 2026; Lee et al., 2026), with hierarchical features and efficient query evaluation (Qin et al., 2024; Li et al., 2025). These representations typically transfer image-language knowledge through region-level observations collected across multiple views.

However, region-to-element transfer introduces an inherent ambiguity: a region descriptor may cover neighboring objects, background, or only part of a target, and thus need not provide equally reliable evidence for every associated 3D element. Repeated views may reinforce the same mixed or incomplete association. Visibility establishes geometric plausibility, but cannot determine how strongly a region descriptor should contribute to each element.

Existing language-field and instance-aware methods provide semantic and object-level cues (Zhou et al., 2024; Shi et al., 2024; Wu et al., 2024). COS3D, for example, couples instance and language fields through feature mapping and adaptive prompting (Zhu et al., 2025). Yet these methods mainly construct or align representations, leaving individual region-to-element reliability insufficiently modeled. This motivates our question: *how should mixed, incomplete, or spatially imprecise multiview semantic evidence be allocated to an individual 3D element?* Geometry, local region structure, and instance membership offer complementary cues for estimating this reliability.

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.

To address this problem, GAFF models observation reliability before integrating multiview evidence into an existing scene representation. Geometric admissibility selects plausible associations; depth consistency, local region agreement, and instance compatibility then determine observation-specific weights. An anchored update combines this evidence with the initial semantics, preserving their contribution while incorporating additional observations. The refinement requires no additional gradient-based optimization.

Reliable semantic association alone, however, does not guarantee accurate query-mask extent. Multilevel language fields can produce semantically strong masks that omit target parts or include surrounding structures. We therefore introduce Object-Guided Mask Refinement (OMR), which uses object candidates from instance-aware scene structure to guide multilevel seed selection and subsequent mask editing. Object-level evidence determines whether candidate regions should be retained, expanded, trimmed, or replaced.

Together, the two stages address consecutive ambiguities: GAFF fusion improves *which semantic evidence should be assigned to each 3D element*, whereas OMR improves *which spatial regions should be assigned to the language query*. They connect semantic reliability with accurate spatial extent without additional scene-specific optimization.

We evaluate both the complete framework and its mechanisms. The full method achieves 68.45% mIoU on LERF-OVS and 94.88% on 3D-OVS. On ten ScanNetV2 scenes, fusion with probability-based readout, spatial averaging, and class-prior calibration, without OMR, achieves 38.28% mIoU over 19 classes. Mass-matched uniform controls preserve accepted observations, total observation mass, and effective anchor proportion, isolating relative weighting from evidence magnitude. Full weighting improves 19-class mIoU by 0.95 percentage points over this control, while observation diagnostics show a larger average weight share for category-correct evidence. LERF-OVS component ablations further demonstrate the complementary benefits of fusion and OMR.

Our main contributions are summarized as follows:

- We introduce GAFF, a refinement framework that uses geometric and instance structure to improve semantic evidence allocation and query-mask prediction without additional gradient-based optimization.
- We combine geometrically admissible, observation-weighted fusion and semantic anchoring with OMR's object-informed seed selection and extent editing to address evidence and spatial-extent ambiguity.
- We evaluate the framework through mass-matched controls, observation-level diagnostics, and component ablations, demonstrating the benefits of relative evidence weighting and complementary mask refinement.

2 Related Work

2.1 Open-vocabulary 3D representations

LERF and 3D-OVS embed vision-language features into radiance fields for text-conditioned localization and segmentation (Kerr et al., 2023; Liu et al., 2023). Feature-3DGS and LEGaussians transfer foundation-model features to Gaussian representations through distillation and quantization (Zhou et al., 2024; Shi et al., 2024), while LangSplat and LangSplatV2 provide hierarchical language features and efficient sparse dictionary representations (Qin et al., 2024; Li et al., 2025). Other approaches explore external semantic indexing (Ding et al., 2026), spatially aware cross-modal alignment and training-free voxel grouping (Liu et al., 2026; Huang et al., 2026), or online feed-forward scene understanding (Lee et al., 2026). These advances improve representation and retrieval; our focus is how to integrate region-level evidence when its correspondence to individual 3D elements is ambiguous.

2.2 Multi-view feature integration

Dr. Splat, econSG, and GeoPurify improve semantic transfer through pixel-ray associations, region-level cross-view aggregation, and geometric distillation, respectively (Kim et al., 2025; Zhang & Lee, 2025; Dou et al., 2026). Instance-aware approaches introduce scale-aware affinities or object-

consistent language representations (Cen et al., 2025; Wu et al., 2024; Piekenbrinck et al., 2025). OpenInsGaussian combines context-aware features with cross-view fusion, while OVSeg3R constrains 3D superpoints using 2D instance masks (Huang et al., 2025; Li et al., 2026). Related systems extend object-aware semantics to online panoptic and open-set mapping (Zhai et al., 2026; Yan et al., 2026). COS3D couples instance and language fields through feature mapping and adaptive prompting (Zhu et al., 2025). These approaches primarily construct or align representations, leaving the contribution of each mixed or incomplete region observation insufficiently modeled. Building on COS3D's representation components, GAFF uses geometric admissibility, local region structure, and instance compatibility to allocate observation-specific evidence before fusion.

2.3 Query-conditioned spatial localization and segmentation

ReferSplat uses position-aware cross-modal interaction for referring segmentation (He et al., 2025). ReLaGS and GeoCGA exploit scene-graph reasoning and semantic-spatial graph alignment, respectively (Xie et al., 2026; Tao et al., 2026), while LangRef3DGS studies referential segmentation from partial observations (Ye et al., 2026). Such spatial reasoning helps identify query-relevant targets, but semantic relevance alone does not determine their complete extent: multilevel masks may omit object parts or include nearby regions. OMR addresses this complementary problem through object-guided multilevel seed selection and extent editing.

3 Method

3.1 Problem formulation

Region-to-element evidence allocation is ambiguous: a region descriptor can span neighboring objects or cover only part of a target, so its contribution need not be equally useful to every associated 3D element. We distinguish two decisions: *which observations* may contribute to an element, and *how much* each accepted observation contributes. GAFF denotes the framework. Its fusion stage addresses these decisions; OMR then uses the updated language responses and object candidates to refine query-mask extent (Figure 1).

Let $\mathcal{P} = \{\mathbf{x}_p\}_{p=1}^{N}$ denote spatial carriers with initial semantic vectors $\mathbf{z}_p$ and instance assignments i_p. A carrier is a Gaussian for feature rendering or a reconstructed point for point classification. An observation $o = (v, r)$ supplies a semantic vector $\mathbf{u}_o$ for region r in camera v. Binary admissibility a_{po} defines the accepted set $\mathcal{O}(p) = \{o : a_{po} = 1\}$. Within this set, nonnegative weights w_{po} define total observation mass $W_p = \sum_o w_{po}$ and, for $W_p > 0$, relative contributions $\alpha_{po} = w_{po}/W_p$. For a finite anchor with mass $\lambda > 0$, fusion takes the form

$$\mathbf{z}'_p = \gamma_p \mathbf{z}_p + (1 - \gamma_p) \sum_{o \in \mathcal{O}(p)} \alpha_{po} \mathbf{u}_o, \qquad \gamma_p = \frac{\lambda}{\lambda + W_p}. \tag{1}$$

Admissibility determines participation, α_{po} allocates evidence among accepted observations, and γ_p sets the anchor balance. Our central hypothesis is that informed allocation improves semantic prediction over uniform allocation even when $\mathcal{O}(p)$, W_p, and hence γ_p, are fixed.

Inputs comprise an RGB Gaussian reconstruction, LangSplatV2 language features, and COS3D instance representations (Kerbl et al., 2023; Li et al., 2025; Zhu et al., 2025); SAM and CLIP supply regions and descriptors (Kirillov et al., 2023; Radford et al., 2021). Fusion updates semantics with geometry and upstream parameters fixed. OMR selects and edits masks rendered from the updated language features. The point branch instead allocates class-score or probability evidence without OMR.

3.2 GAFF fusion stage

The GAFF fusion stage implements evidence allocation through geometric admissibility, observation-specific weighting, and semantic anchoring.

Geometric admissibility: which observations. A carrier projects into view v as $\mathbf{a}_{pv} = \pi(\mathbf{K}_v(\mathbf{R}_v\mathbf{x}_p + \mathbf{t}_v))$, where π converts homogeneous coordinates to pixels. We retain positive-depth projections with valid image coordinates and region IDs. A local point z-buffer provides a

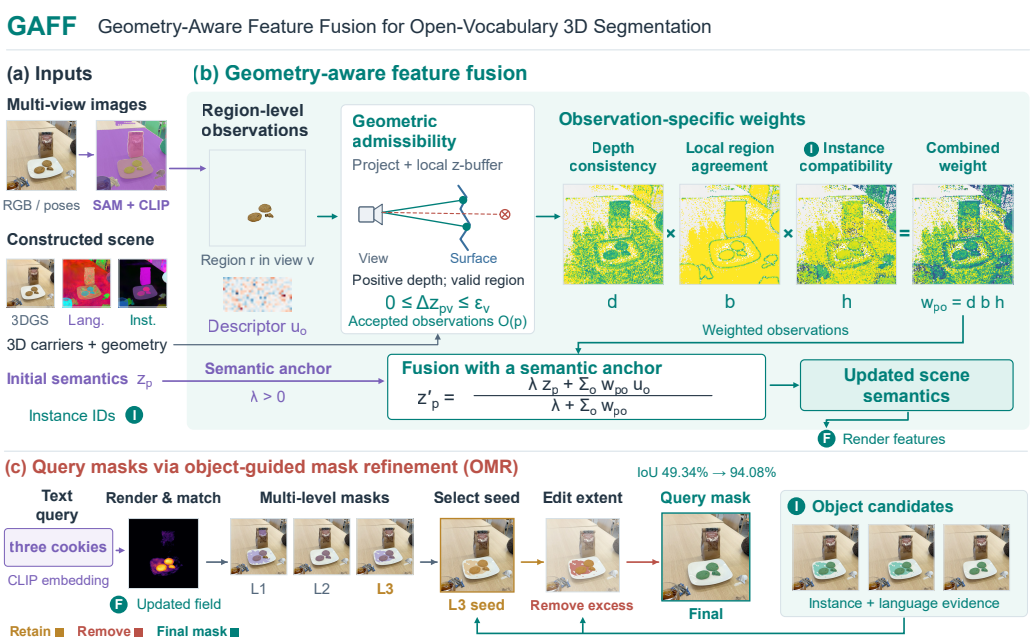

Figure 1: The GAFF framework: (a) inputs, (b) GAFF fusion stage, and (c) OMR. Geometric admissibility selects region-to-element associations; depth, local-region, and instance cues weight accepted evidence before semantic anchoring. OMR subsequently adjusts query-mask extent. The three-cookies example illustrates this downstream step: editing increases seed IoU from 49.34% to 94.08%.

reference surface, and a binding is accepted when its depth residual satisfies $0 \le \Delta z_{pv} \le \epsilon_v$. The accepted observations form $\mathcal{O}(p)$, with $a_{po} = 0$ for rejected associations. This hard geometric test approximates visibility for association; it does not establish semantic reliability. Feature rendering separately accounts for accumulated transmittance.

Observation weighting: how much contribution. For each accepted observation, we combine depth consistency, local region agreement, and instance compatibility:

$$w_{po} = d_{pv} b_{pv} h_{pv}, \qquad d_{pv} = \exp(-\Delta z_{pv}/\delta_v), \qquad h_{pv} = \eta + (1-\eta)\frac{n_{i_p vr}}{\sum_{r'} n_{i_p vr'}}. \tag{2}$$

We set $\epsilon_v = \delta_v$, tying the admissibility tolerance to the depth-decay scale; Appendix C specifies its dataset-dependent value. The count n_{ivr} records visible valid bindings from instance i to region r in view v. The instance term measures the fraction of visible instance elements assigned to the region, with floor η; unknown instance assignments use $h_{pv} = 1$. This is an instance-coverage fraction, not a measure of region purity: a region covering several instances can receive high compatibility for each. For the region map R_v, let $\mathcal{D} = \{(\pm k, 0), (0, \pm k) : k = 1, \dots, B_0\}$. The indicator $\chi_{pv}(\mathbf{d})$ is one when the offset stays inside the image and retains the projected region ID. The local-region term is

$$b_{pv} = \beta + \frac{1-\beta}{|\mathcal{D}|} \sum_{\mathbf{d} \in \mathcal{D}} \chi_{pv}(\mathbf{d}). \tag{3}$$

The factors describe surface agreement, consistency within a local region neighborhood, and coverage of the assigned instance. Their product lowers a contribution when any factor is small. The floors in b_{pv} and h_{pv} retain a residual contribution. The weights scale the whole region descriptor; they do not separate object-specific semantics within a mixed descriptor.

Semantic anchoring. The initial semantic vector provides a reference against which additional observations refine the field:

$$\mathbf{z}'_p = \frac{\lambda \mathbf{z}_p + \sum_{o \in \mathcal{O}(p)} w_{po} \mathbf{u}_o}{\lambda + \sum_{o \in \mathcal{O}(p)} w_{po}}, \qquad \lambda > 0. \tag{4}$$

For a finite anchor and $W_p > 0$, this is Eq. 1. Changing weights can change both α_{po} and γ_p; fixing λ alone does not fix the effective anchor proportion. Our mass-matched uniform control sets $w_{po}^{\mathrm{MM}} = W_p / K_p$, where $K_p = |\mathcal{O}(p)|$, using Full's observation mass. It changes only relative allocation to $\alpha_{po}^{\mathrm{MM}} = 1/K_p$, preserving participation and anchor balance. Both retain the same fallback rules: zero-observation-mass carriers keep their initial vectors; nonfinite anchors have zero initial mass (Appendix C). Language features are fused in decoded teacher spaces before rendering, normalization, and text matching. Point controls use cosine scores; the optimized branch uses probabilities, spatial averaging, and prior calibration (Appendix G). Section 4.3 tests relative allocation under the matched control.

3.3 OMR AFTER FUSION

After evidence allocation, query-mask extent remains a separate ambiguity: language relevance need not identify the full target or exclude its surroundings. Within the GAFF framework, OMR complements the fusion stage through object-informed multilevel seed selection and candidate-guided extent editing. It uses object structure to retain or recover target regions and remove extraneous responses from the output mask.

Multilevel query evidence. At level l, the language field $\mathbf{e}_l$ is rendered and normalized as $\mathbf{f}_{v,l}(\mathbf{a}) = \mathrm{norm}(\mathcal{R}_v[\mathbf{e}_l](\mathbf{a}))$. For normalized query embedding $\mathbf{t}_q$, relevance is measured relative to a fixed negative bank $\mathcal{N} = \{\mathrm{object}, \mathrm{things}, \mathrm{stuff}, \mathrm{texture}\}$:

$$\rho_{v,l}(\mathbf{a}, q) = \sigma\left(\kappa_\rho\left[\langle \mathbf{f}_{v,l}(\mathbf{a}), \mathbf{t}_q \rangle - \max_{n \in \mathcal{N}} \langle \mathbf{f}_{v,l}(\mathbf{a}), \mathbf{t}_n \rangle\right]\right). \tag{5}$$

Smoothing, contrast normalization, and thresholding yield masks M_l. Confidence c_l is measured before contrast normalization; the language-only seed uses $r = \arg\max_l c_l$. This selects response strength, not necessarily suitable extent. Decoding details are in Appendix A.

Object-informed seed selection. Instance-language associations and projected object crops supply candidate regions and semantic descriptors. Crop descriptors are pooled with area weights as $\bar{\mathbf{v}}_j = \mathrm{norm}(\sum_v a_{jv} \mathbf{v}_{jv})$. Candidate overlap, coverage, compactness, and language confidence determine the selected level l^* and seed $B = M_{l^*}$. The seed's structure also affects subsequent editing gates. Selection is therefore assessed through the final edited prediction, not an assumed improvement in standalone seed IoU (Section 4.5).

Candidate-informed extent editing. The crop-based route ranks proposals by semantic relevance and object structure, then applies gates for replacement or local patching. A multilevel coverage route instead combines language masks and instance candidates through coverage and morphology rules. The ranking equation, dataset-specific gates, and fallback behavior are specified in Appendix D.

Let $\mathcal{A}_q$ and $\mathcal{D}_q$ denote the sets of pixels accepted for addition to and removal from the seed mask B, respectively. The output is then

$$Y = (B \setminus \mathcal{D}_q) \cup \mathcal{A}_q. \tag{6}$$

Empty edit sets retain the seed. This downstream mask operation is evaluated separately from semantic fusion in Section 4.

Table 1: GAFF framework: LERF-OVS scene-level segmentation results (%). †LangSplatV2 mIoU is quoted from its original paper (Li et al., 2025); its localization accuracy uses a different definition and is omitted from the mAcc columns. Sections 4.1–4.2 specify the scope of external comparisons.

Method	figurines		teatime		ramen		waldo_kitchen		mean	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
LangSplat	10.16	8.93	11.38	20.34	7.92	11.27	9.18	9.09	9.66	12.41
LEGaussians	17.99	23.21	19.27	27.12	15.79	26.76	11.78	18.18	16.21	23.82
OpenGaussian	39.29	55.36	60.44	76.27	31.01	42.25	22.70	31.82	38.36	51.43
OpenInsGaussian	53.75	76.79	58.55	77.97	26.38	43.66	31.78	50.00	42.62	62.11
Dr. Splat	53.36	80.36	57.20	76.27	24.70	35.21	39.07	63.64	43.58	63.87
COS3D	60.03	82.14	65.07	91.53	35.86	46.48	42.10	68.18	50.76	72.08
OpenSplat3D	60.71	85.71	73.27	88.14	49.20	76.06	55.63	77.27	59.70	81.79
LangSplatV2†	56.40	–	72.20	–	51.80	–	59.10	–	59.90	–
LangRef3DGS	62.52	86.65	73.41	87.81	49.73	77.50	57.11	77.68	60.69	82.41
GAFF	**63.87**	**87.50**	**77.98**	**93.22**	**68.02**	**90.14**	**63.91**	**90.91**	**68.45**	**90.44**



Figure 2: LERF-OVS query-mask comparisons. The displayed examples show reduced leakage around the dall-e brand and toy queries, retention of thin knife regions, and a more complete sink mask. They illustrate extent differences between the complete GAFF framework and LangSplatV2.

4 Experiments

4.1 Evaluation setup

We evaluate the proposed framework from both system-level and mechanism-level perspectives. We first assess complete-system performance on LERF-OVS, followed by 3D-OVS and ScanNetV2, which together cover rendered query-mask prediction and point-level semantic classification. We then conduct controlled ablations and mechanism analyses to isolate the contribution of relative observation weighting and to quantify the additional gains introduced by OMR after feature fusion.

For LERF-OVS and 3D-OVS, we follow the standard rendered query-mask evaluation protocols and report scene-averaged results (Kerr et al., 2023; Liu et al., 2023; Qin et al., 2024; Zhu et al., 2025). For ScanNetV2, we adopt the ten-scene setting and the 19-, 15-, and 10-class category splits used by OpenGaussian, and evaluate all valid native ground-truth points without applying OMR (Dai et al., 2017; Wu et al., 2024). All controlled comparisons reuse fixed scene representations so that observed differences can be attributed to the evaluated refinement mechanisms rather than representation retraining. Detailed settings for training, feature association, fusion, and evaluation are provided in Appendices E, C, and F.

4.2 Overall performance across output settings

LERF-OVS. Table 1 reports the scene-level performance of the complete framework. GAFF achieves 68.45% mIoU and obtains higher reported mIoU and mAcc than LangRef3DGS across

Table 2: GAFF framework: 3D-OVS mIoU (%). External baselines use the reported values in the cited papers.

Method	Bed	Bench	Room	Sofa	Lawn	Average
3D-OVS	89.5	89.3	92.8	74.0	88.2	86.8
Feature-3DGS	83.5	90.7	84.7	86.9	93.4	87.8
LEGaussians	84.9	91.1	86.0	87.8	92.5	88.5
LangSplat	92.5	94.2	94.1	90.0	96.1	93.4
LangSplatV2	93.0	94.9	**96.1**	92.3	96.6	94.6
ReferSplat	90.2	93.8	94.1	90.8	95.5	92.9
GeoCGA	90.1	**95.4**	93.7	**93.4**	96.0	93.7
GAFF	**93.2**	95.3	95.7	92.9	**97.3**	**94.9**

Table 3: GAFF framework on ScanNetV2 (%), with optimized readout and no OMR. The blocks use different evaluation settings.

Method	19 classes		15 classes		10 classes	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
OpenGaussian scene/category sets						
LangSplat	3.78	9.11	5.35	13.20	8.40	22.06
LEGaussians	3.84	10.87	9.01	22.22	12.82	28.62
OpenGaussian	24.73	41.54	30.13	48.25	38.29	55.19
COS3D	32.47	49.05	35.95	54.35	44.32	63.66
OpenInsGaussian	37.50	54.38	**38.14**	55.30	**51.42**	69.15
GAFF	**38.28**	**62.34**	38.10	**61.54**	47.20	**71.29**
Other settings: reference only						
Dr. Splat	28.38	55.85	31.84	60.58	39.21	66.66
EmbodiedSplat	33.12	59.03	37.90	66.82	45.56	75.13

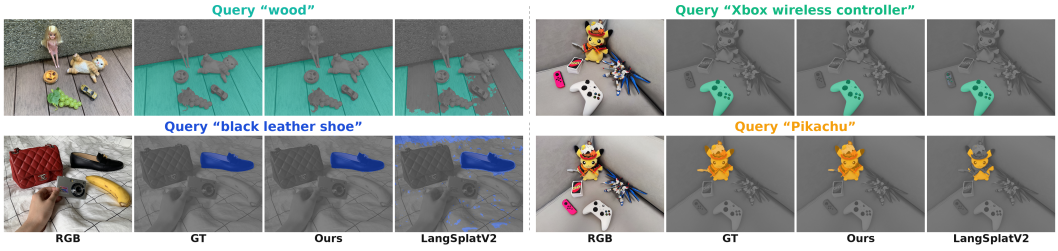

Figure 3: 3D-OVS query-mask comparisons. The wood example illustrates more complete coverage and the shoe example less surrounding response, while the controller mask is retained. These selected examples complement the aggregate results.

all four scenes. To examine whether this improvement depends on annotation-processing conventions, we further re-score the same predictions using the polygon rasterization protocol adopted by LangSplatV2. The resulting mIoU remains nearly unchanged, indicating that the observed performance is robust to these two annotation-processing choices (Appendix H.3). These results demonstrate that the proposed refinement framework consistently improves open-vocabulary query-mask prediction under the LERF-OVS evaluation setting.

3D-OVS. Table 2 evaluates the same framework on five additional scenes from 3D-OVS. GAFF achieves a mean mIoU of 94.88%, exceeding LangSplatV2, the strongest baseline listed in the comparison. The qualitative results further illustrate that object-guided refinement can recover missing target regions while suppressing responses from surrounding structures. These observations are consistent with the intended role of OMR in correcting spatial-extent ambiguity after semantic fusion. Section 4.5 further isolates the effects of multilevel seed selection and extent editing through controlled component studies.

ScanNetV2. Table 3 evaluates native-point labels without OMR for 19, 15, and 10 classes. The point-prediction configuration combines fusion with probability readout, spatial averaging, and class-prior calibration (Appendix G); fixed-readout fusion controls follow in Section 4.3.

The upper block uses OpenGaussian's scenes and category sets, retaining all valid GT points. Dr. Splat and EmbodiedSplat (Kim et al., 2025; Lee et al., 2026) use different settings and are listed separately. Baseline values come from external reports; Appendix F details the evaluation domains.

4.3 Semantic prediction gains without mask editing

Hypothesis and control. We hypothesize that relative evidence allocation improves semantic prediction at fixed participation and anchor balance. To test this, we compare Full weighting with mass-matched Uniform. Both use the same geometric admissibility, accepted associations, features,

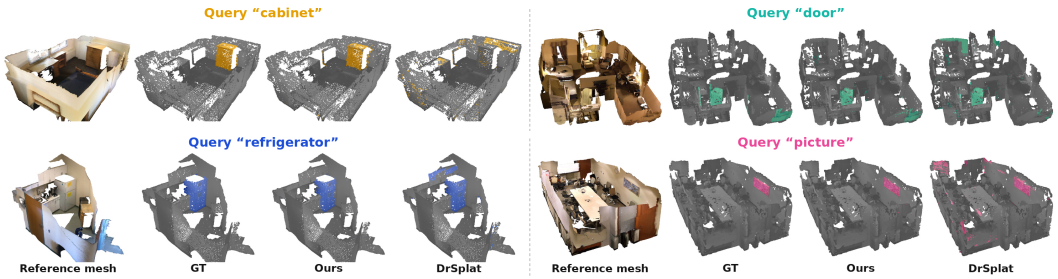

Figure 4: ScanNet point-label comparisons under the displayed settings. The Ours column shows the GAFF framework with optimized point readout. The evaluation-domain distinctions in Table 3 also apply to the interpretation of these examples.

Table 4: Relative evidence allocation without OMR (mIoU, %). All fusion rows share accepted associations and cosine readout. The primary block fixes per-point observation mass and effective anchor proportion.

Configuration	19 classes	15 classes	10 classes
Primary control: matched observation mass			
Uniform / matched mass	36.7398	36.0969	44.3463
GAFF fusion stage (*dbh*)	**37.6899**	**37.0852**	**45.6601**
Additional references			
Original	33.4675	32.8591	41.1075
Uniform	36.7840	36.1381	44.4393
Soft geometry (*db*)	36.8816	36.2330	44.6027

per-point observation mass, and cosine readout; only relative weights differ. OMR, probability readout, spatial averaging, and prior calibration are disabled. All valid native GT points are scored, including those without accepted observations, with equal-scene averaging. Predictions are frozen before GT is used for evaluation or diagnosis.

Relative allocation beyond mass effects. Full outperforms mass-matched Uniform under all three class settings, with improvements in most scenes (Table 4). These comparisons support the hypothesis that relative allocation improves semantic prediction: with total observation mass and effective anchor proportion held fixed, redistributing weight among accepted observations yields consistent gains across the three category sets.

Observation-level diagnostics. We further hypothesize that the GAFF fusion stage allocates more weight to category-correct evidence. To test this, we compare its unweighted observation share with its share of total fusion weight. A category-correct observation has an unfused region cosine argmax matching the point's GT class. For the 19-class setting, Full weighting increases the average weight share assigned to category-correct observations, with this trend holding across most scenes. This diagnostic averages within scenes and then equally across scenes, using valid GT points with finite observations and positive mass. The increased weight share supports this hypothesis: weighting preferentially strengthens category-correct evidence within the accepted observation set. Together with the mass-matched prediction gains, this provides evidence for the effectiveness of element-specific relative weighting in semantic fusion.

Appendix H.1 provides factor controls and scene-level analyses. The mass-matched comparison isolates the effect of relative weighting, while the observation-level diagnostic supports its allocation of weight toward category-correct evidence.

4.4 Fusion and Full-Framework Results

We hypothesize that object-guided refinement improves fused predictions by addressing remaining extent ambiguity. To test this, Table 5 varies only fusion and OMR with inputs, thresholds, and evaluation fixed. Both Fusion only and OMR only improve over Baseline, and combining them out-

Table 5: LERF-OVS component configurations (mIoU, %). F: fusion; P: OMR; GAFF: full framework.

Configuration	F	P	mIoU
Baseline			57.8970
Fusion only	✓		59.1284
OMR only		✓	64.5000
GAFF	✓	✓	**68.4450**

Table 6: OMR stage ablations (mIoU, %). GAFF denotes the full framework; other stages and native inputs are retained.

Configuration	3D-OVS	LERF-OVS
GAFF	**94.8776**	**68.4450**
w/o object-informed selection	89.3494	67.9195
w/o object-guided editing	91.6460	54.7334

performs either component alone. These results support the hypothesis that OMR further improves fused predictions by refining their spatial extent. They demonstrate the complementary roles of the two stages within the GAFF framework: fusion refines the semantic evidence used for prediction, and OMR subsequently selects and adjusts the resulting query-mask extent.

4.5 Seed selection and extent editing ablations

We hypothesize that object-informed seed selection and extent editing jointly improve the final mask. We test this by removing either stage while retaining native inputs and thresholds. Without selection, the editor uses the language-only seed; without editing, the selected seed is returned. The latter retains selection, whereas matched Fusion only disables OMR entirely (Appendix D.5).

Object-informed seed selection improves the final edited masks on both datasets, while extent editing further improves predictions based on the selected seeds. Their combination achieves the best mask accuracy among the tested OMR configurations (Table 6), demonstrating the effectiveness of jointly selecting a seed using object structure and adjusting its spatial extent.

These results support the hypothesis that selection and editing work together to improve the final output: object-informed selection supplies the seed for candidate-guided editing, which refines target coverage and removes excess response. This validates OMR as a downstream complement to semantic fusion for resolving query-mask extent ambiguity.

Figure 1 illustrates excess-region removal; Appendices B and D.5 provide retention/completion examples and per-scene ablation details.

5 Discussion and Limitations

The results support relative evidence weighting beyond geometric admissibility, with OMR providing complementary refinement of query-mask extent.

Fusion depends on reliable region semantics, geometric associations, and instance partitions. Repeated incorrect observations can reinforce errors (scene0140, Appendix H.2), motivating better observation filtering and handling of regions that mix multiple objects.

Our evaluation covers public benchmark scenes. Extending GAFF to dynamic scenes and incremental updates remains challenging, as evidence allocation and object structure must remain consistent with evolving observations.

6 Conclusion

Open-vocabulary 3D semantic integration requires both admissible observations and explicit modeling of their contributions to each element. The GAFF fusion stage combines geometric admissibility, observation-specific weighting, and semantic anchoring. Mass-matched controls establish gains from relative weighting at fixed observation mass and anchor balance, supported by observation-level diagnostics. OMR then uses object-guided seed selection and editing to address the remaining mask-extent ambiguity. The results support evidence allocation as the core semantic integration step and object-guided extent refinement as its subsequent complement.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. The code is available in the anonymous repository at `https://anonymous.4open.science/r/Geometry-Aware-Feature-Fusion-for-Open-Vocabulary-3D-Segmentation` to facilitate replication and verification. The experimental setup, including the detailed method pipeline, training procedures, and hardware specifications, is described in the main text and appendices.

All datasets used in this work, including LERF-OVS, 3D-OVS, and ScanNetV2, are publicly available to facilitate consistent and reproducible evaluation. We believe that these measures will enable other researchers to reproduce our work and advance research in open-vocabulary 3D segmentation.

AI USE STATEMENT

Large language models were used to assist with polishing, translation, editing, and consistency checks of author-provided text, as well as literature retrieval and discovery of related work. They also assisted with rapid feasibility checks of research ideas, following the authors' explicitly specified ideas and processing procedures. The authors formulated the research ideas, designed and implemented the method, and conducted the experiments. AI assistance was used to facilitate these author-directed feasibility checks, rather than to independently develop or implement the proposed method. The authors remain responsible for verifying the manuscript, references, experimental results, and scientific conclusions.

REFERENCES

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment Any 3D Gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

Jiayu Ding, Xinpeng Liu, Zhiyi Pan, Shiqiang Long, and Ge Li. ExtrinSplat: Decoupling Geometry and Semantics for Open-Vocabulary Understanding in 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31019–31028, June 2026.

Weijia Dou, Xu Zhang, Yi Bin, Jian Liu, Bo Peng, Guoqing Wang, Yang Yang, and Heng Tao Shen. GeoPurify: A Data-Efficient Geometric Distillation Framework for Open-Vocabulary 3D Segmentation. In *International Conference on Learning Representations*, pp. 1557–1578, 2026.

Shuting He, Guangquan Jie, Changshuo Wang, Yun Zhou, Shuming Hu, Guanbin Li, and Henghui Ding. ReferSplat: Referring Segmentation in 3D Gaussian Splatting. In *International Conference on Machine Learning*, 2025.

Sheng-Yu Huang, Jaesung Choe, Yu-Chiang Frank Wang, and Cheng Sun. OpenVoxel: Training-Free Grouping and Captioning Voxels for Open-Vocabulary 3D Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16734–16745, June 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (Findings)*, pp. 6400–6409, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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 (CVPR)*, pp. 14137–14146, June 2025.

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.

Seungjun Lee, Zihan Wang, Yunsong Wang, and Gim Hee Lee. EmbodiedSplat: Online Feed-Forward Semantic 3DGS for Open-Vocabulary 3D Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23774–23784, 2026.

Hongyang Li, Jinyuan Qu, and Lei Zhang. OVSeg3R: Learn Open-vocabulary Instance Segmentation from 2D via 3D Reconstruction. In *International Conference on Learning Representations*, pp. 121352–121371, 2026.

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. *Advances in Neural Information Processing Systems*, 2025.

Jiarun Liu, Qifeng Chen, Yiru Zhao, Minghua Liu, Baorui Ma, and Sheng Yang. TIGaussian: Disentangle Gaussians for Spatial-Awared Text-Image-3D Alignment. In *International Conference on Learning Representations*, pp. 928–940, 2026.

Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly Supervised 3D Open-vocabulary Segmentation. In *Advances in Neural Information Processing Systems*, 2023.

Jens Piekenbrinck, Christian Schmidt, Alexander Hermans, Narunas Vaskevicius, Timm Linder, and Bastian Leibe. OpenSplat3D: Open-Vocabulary 3D Instance Segmentation using Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5285–5294, 2025.

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*, 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 International Conference on Machine Learning*, volume 139, pp. 8748–8763, 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.

Yuwen Tao, Kanglei Zhou, Chang Li, and Liyuan Wang. Geometry-Aware Cross-Modal Graph Alignment for Referring Segmentation in 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20433–20442, June 2026.

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. *Advances in Neural Information Processing Systems*, 2024.

Yaxu Xie, Abdalla Arafa, Alireza Javanmardi, Christen Millerdurai, Jia Cheng Hu, Shaoxiang Wang, Alain Pagani, and Didier Stricker. ReLaGS: Relational Language Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23826–23836, June 2026.

Yucheng Yan, Chen Liang, Wenguan Wang, and Yi Yang. Open-Set Semantic Gaussian Splatting SLAM with Expandable Representation. In *International Conference on Learning Representations*, pp. 91986–92013, 2026.

Xulun Ye, Qin Zhang, and Kun Zhou. LangRef3DGS: Natural Language-Guided 3D Referential Segmentation from Partial Observations via 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38595–38605, June 2026.

Hongjia Zhai, Qi Zhang, Xiaokun Pan, Xiyu Zhang, Yitong Dong, Huaqi Zhang, Dan Xu, and Guofeng Zhang. OnlinePG: Online Open-Vocabulary Panoptic Mapping with 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33269–33279, June 2026.

Can Zhang and Gim H Lee. econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians. In *International Conference on Learning Representations*, pp. 68368–68385, 2025.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024.

Runsong Zhu, Ka-Hei Hui, Zhengzhe Liu, Qianyi Wu, Weiliang Tang, Shi Qiu, Pheng-Ann Heng, and Chi-Wing Fu. COS3D: Collaborative Open-Vocabulary 3D Segmentation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

GEOMETRY-AWARE FEATURE FUSION FOR OPEN-VOCABULARY 3D SEGMENTATION SUPPLEMENTARY MATERIAL

The GAFF fusion stage uses geometry and instance structure to guide the integration of region-level language observations into existing scene semantics. OMR then addresses the remaining query-mask extent decision. Appendices A–D specify the representations, fusion, and complete prediction algorithms; E–F give training and evaluation; G specifies the optimized ScanNet readout; H reports matched experiments and extended results. The term *fusion* denotes the GAFF fusion stage, while GAFF denotes the framework.

A LANGUAGE FIELDS, RENDERING, AND QUERY READOUT

Gaussian feature rendering. A 3D Gaussian scene comprises primitives with centers $\mathbf{x}_p$, covariances $\mathbf{\Sigma}_p$, opacities o_p, and appearance attributes. Given camera v, each primitive projects to a 2D footprint whose opacity contribution at pixel $\mathbf{a}$ is $\alpha_{pv}(\mathbf{a})$. The same rasterization procedure renders a semantic vector $\mathbf{e}_p$ attached to each primitive:

$$\mathcal{R}_v[\mathbf{e}](\mathbf{a}) = \sum_{p \in \mathcal{G}_v(\mathbf{a})} T_{pv}(\mathbf{a}) \alpha_{pv}(\mathbf{a}) \mathbf{e}_p, \qquad T_{pv}(\mathbf{a}) = \prod_{t \prec p} \big(1 - \alpha_{tv}(\mathbf{a})\big). \tag{S1}$$

Here $\mathcal{G}_v(\mathbf{a})$ contains contributing Gaussians in front-to-back order, and $t \prec p$ denotes a primitive preceding p along the ray. This provides a common interface for projecting language features and instance membership into the image plane.

Multilevel language decoding. The core OVS profiles use LangSplatV2 multilevel language assets. At level l, a dictionary $\mathbf{D}_l$ stores language vectors, and each Gaussian has coefficient logits $\mathbf{a}_{p,l}$. Its decoded feature is

$$\mathbf{e}_{p,l} = \sum_{d \in \mathcal{K}_{p,l}} \frac{\exp(a_{p,l,d})}{\sum_{d' \in \mathcal{K}_{p,l}} \exp(a_{p,l,d'})} \mathbf{D}_{l,d}, \tag{S2}$$

where $\mathcal{K}_{p,l}$ indexes the top-k logits and $\mathbf{D}_{l,d}$ is the d-th dictionary vector. Each level is decoded in its own 512-dimensional teacher space before feature fusion. A Gaussian's sparse coefficients from one dictionary are never interpreted in another dictionary. The decoder applies softmax over the selected logits and computes a weighted sum of the corresponding dictionary vectors, rather than an unweighted average.

Rendering and query readout. The rendering comparisons keep geometry, opacity, cameras, and the `gsplat` feature renderer fixed. Feature vectors are rendered before normalization. We use $\mathrm{norm}(\mathbf{e}) = \mathbf{e}/\|\mathbf{e}\|_2$ for nonzero-vector normalization and $\mathbf{t}_q$ for the normalized embedding of query q. For a rendered normalized feature $\mathbf{f}$, query relevance uses the hardest negative:

$$\rho(\mathbf{f}, q) = \min_{n \in \mathcal{N}} \frac{\exp(10\langle \mathbf{f}, \mathbf{t}_q \rangle)}{\exp(10\langle \mathbf{f}, \mathbf{t}_q \rangle) + \exp(10\langle \mathbf{f}, \mathbf{t}_n \rangle)}. \tag{S3}$$

The negative bank is {object, things, stuff, texture}. Let $\mathcal{A}_k$ be a $k \times k$ average filter that excludes out-of-image padding from its denominator. The decoder computes $\tilde{\rho} = (\rho + \mathcal{A}_{29}(\rho))/2$ and the response

$$H = \mathrm{clip}\left(2\frac{\tilde{\rho} - \min \tilde{\rho}}{\max \tilde{\rho} - \min \tilde{\rho} + 10^{-9}} - 1, 0, 1\right), \qquad M = \mathbb{1}[\mathcal{A}_7(\mathbb{1}[H > \tau]) > 0.5]. \tag{S4}$$

The binarization threshold τ follows the dataset-specific settings in Appendix D and is held fixed across component configurations. Extrema are taken within a single query/level image. Level confidence is measured before this response normalization: 3D-OVS averages $\tilde{\rho}$ inside M, whereas LERF uses $\max \tilde{\rho}$. Empty masks receive confidence -1. Using the normalized peak would make levels incomparable because each nonconstant response map would peak at one.

Table S1: Default language-interface settings. The readout has dataset-specific admissible language-reference levels. Component comparisons share the same dataset-specific settings.

Setting	3D-OVS	LERF-OVS
Decoded levels	L1/L2/L3	L1/L2/L3
Admissible readout levels	L2/L3	L1/L2/L3
Decoder top-k	4	4
Mask threshold	0.4	0.25
Language reference confidence	Region mean	Smoothed peak

Confidence, ties, and numerical conventions. For level l, let $A_l = \sum_{\mathbf{a}} M_l(\mathbf{a})$. The two confidence statistics used by the language reference are

$$c_l^{\text{mean}} = \frac{\sum_{\mathbf{a}} M_l(\mathbf{a})\tilde{\rho}_l(\mathbf{a})}{A_l}, \qquad c_l^{\text{peak}} = \max_{\mathbf{a}} \tilde{\rho}_l(\mathbf{a}), \quad A_l > 0. \tag{S5}$$

Both are set to -1 if $A_l = 0$. The reference chooses the maximum over its admissible levels, with ties resolved toward the smaller level number. The native object-informed selector can subsequently choose another available level, including L1 for a 3D-OVS rescue. A constant response produces an empty thresholded mask. Queries are evaluated independently, so masks for different queries may overlap.

Field decoding and rendering settings. For each scene and level, decode the trained language field, optionally apply the fusion accumulation in Appendix C, and render before normalization and query matching. The fused vectors remain in the decoded 512-dimensional teacher space, without codebook projection. Matched comparisons use the same renderer, image resolution, camera scaling, and mask thresholds. Table 5 varies only the fusion and OMR switches on the same upstream representations and evaluation protocol. Retained continuous fields enable new text queries, whose first use may require materializing the multilevel fields.

B Workflow and Functional Terminology

Inputs, outputs, and terminology. A *region observation* is a 2D SAM region with a camera and a CLIP vector. An *instance assignment* is an estimated partition of 3D carriers used for fusion compatibility counts. An *object proposal* is a candidate extent used by OMR; it may depend on the query and the selected seed. These objects serve different operations and need not have a one-to-one correspondence. A *language reference* selects a level from language confidence alone; an *object-informed seed* additionally uses object evidence; a *complete output* includes editing and any task-specific final readout.

The execution order is: construct and freeze representations; associate region evidence; apply the selected field-fusion profile; decode query masks; select and edit a seed with OMR. ScanNet instead converts evidence to class scores for the matched cosine controls, or class probabilities for the optimized profile in Appendix G. Both aggregate on the native point domain without executing OMR.

Complementing the deletion example in main-paper Figure 1, Figures S1 and S2 illustrate seed retention and completion, respectively. Figure S1 illustrates the workflow using the query "sheep" in frame 00025 of the LERF-OVS teatime scene. The ten modules below cover scene representation, the GAFF fusion stage, and OMR. The group numbers identify visualized modules rather than imposing a strictly sequential execution order: object candidates can inform seed selection, while seed-route proposals depend on the selected mask.

00. Scene input. Prepare RGB images, camera parameters, coordinate conventions, and text queries. The overview uses a fixed evaluation frame and aligned crops. Ground-truth masks are used only for evaluation and diagnostic checks.

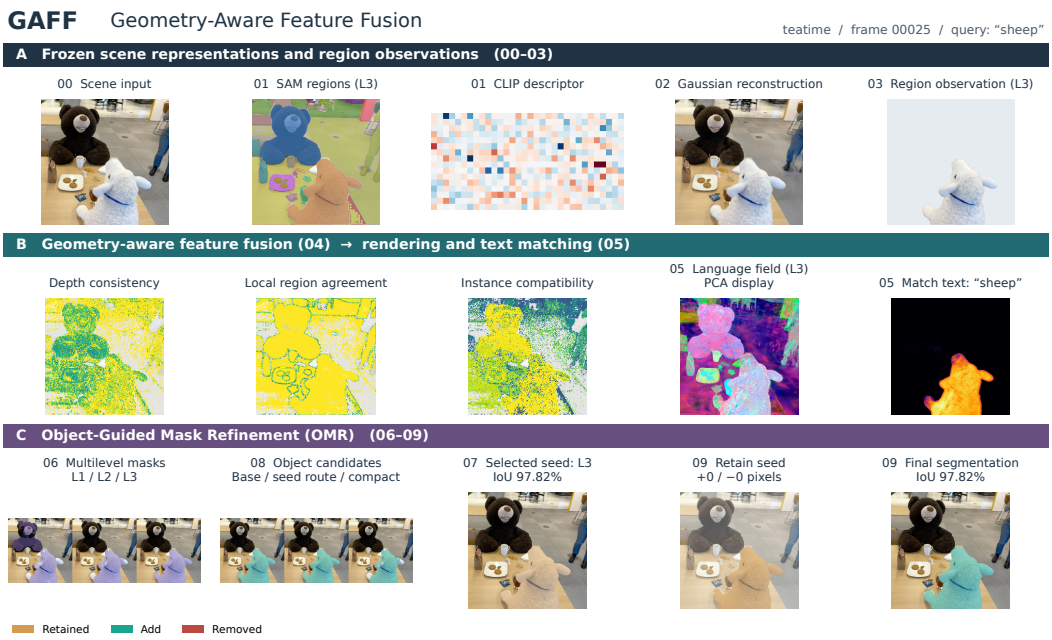

Figure S1: Workflow overview for the teatime "sheep" query. The panels pair scene input, SAM/CLIP features, Gaussian reconstruction, and region observations with fusion weights and the OMR operations of text matching, multilevel mask construction, candidate generation, seed selection, and extent editing. This example retains the L3 seed, with no added or removed pixels, and reaches 97.82% IoU. Fusion-weight panels are evaluation-view diagnostics; the displayed responses, masks, and decisions come from saved full-controller outputs. Feature tiles and PCA colors visualize feature vectors, not spatial probabilities.

01. SAM/CLIP preprocessing. Extract three levels of SAM regions and their 512-dimensional CLIP descriptors. The overview pairs the SAM region map with a descriptor tile; the tile arranges feature components for display and has no image-space interpretation.

02. Scene representation. Reconstruct the RGB Gaussian scene, then initialize and train the multilevel language and auxiliary instance branches from this reconstruction. The corresponding assets include a trained Gaussian RGB rendering, projected ellipses, depth, and language/instance feature visualizations. PCA is used only for display. The newly rendered expected depth differs from the local z-buffer reference used for observation association.

03. Region observations. Associate each image region with its camera and CLIP vector to form the observations used by fusion. The supplied assets show three semantic scales of the same frame, not three different cameras; the overview displays the L3 region observation.

04. GAFF fusion stage. Project scene elements into region maps, establish visible associations, and compute depth consistency, local region agreement, and instance compatibility. Their product weights the region features; semantic anchoring integrates them with the initial semantics in the update of Eq. 4. The three overview panels visualize these factors.

05. Rendering and text matching. Render the language fields, normalize the rendered vectors, and match them to the CLIP text embedding of "sheep" to obtain query responses at each level. The response maps used in this example are saved full-controller outputs; new feature-PCA renderings do not replace those responses.

06. Multilevel mask construction. Decode the L1/L2/L3 response maps into binary masks using the recorded threshold and majority-filter settings. The assets provide the masks, their RGB overlays, and a side-by-side comparison. For this example, decoding the saved responses reproduces all three stored masks pixelwise.

07. Level selection. Compare language confidence and object evidence to choose a seed mask. The recorded L1/L2/L3 confidence values are 0.6161, 0.6962, and 0.7396; the selector

retains L3 in this frame. Selection determines the initial mask for subsequent editing rather than directly committing to an object proposal.

08. Object candidates. Connect instance structure and language through instance–language feature pairs and candidate associations. The candidates include the base-language mask, a proposal derived from the selected seed, and an alternate compact-level proposal, each with its mask, crop features, and scores. Candidate relevance, spatial overlap, and structural scores inform seed selection and the subsequent editing decision; seed-route proposals are constructed from the selected mask.

09. Extent editing and final segmentation. Compare the selected seed and object candidates to retain, add, or remove regions, producing the final mask. In this frame, the selected seed-route candidate coincides with the seed and the controller retains the seed: 92,133 pixels are retained, with zero additions and zero removals. The final mask is identical to the L3 seed and has an independently checked IoU of 97.8179%.

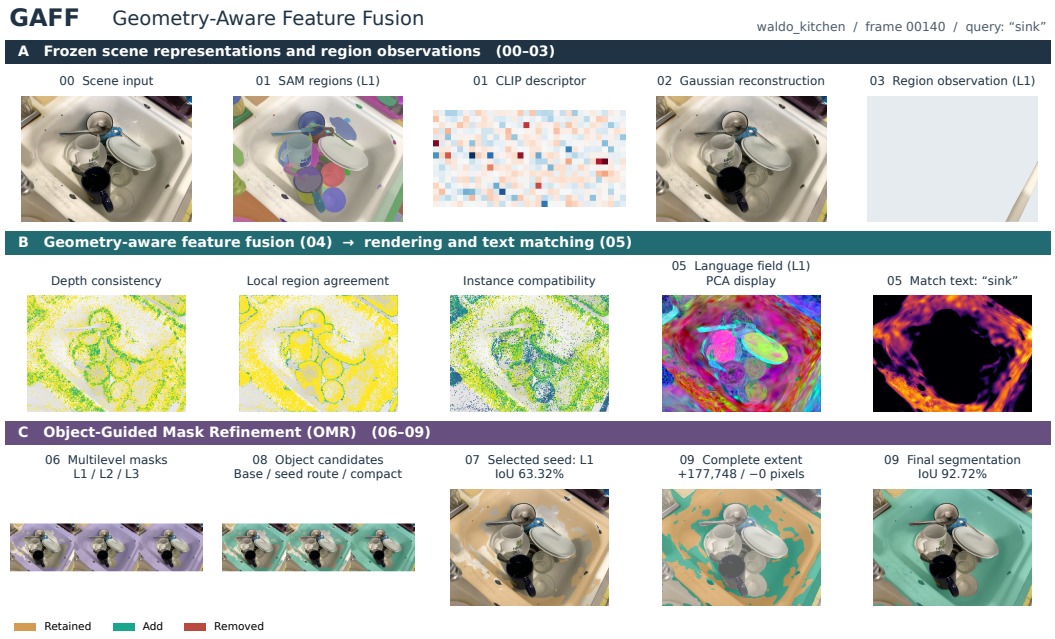

Figure S2: Workflow and object-informed completion for "sink" in frame 00140 of waldo_kitchen. The panels show scene and region representations, diagnostic fusion weights, multilevel query masks, object candidates, the selected L1 seed, the extent edit, and the final segmentation. OMR completes the seed using an accepted L3 candidate. Gold denotes retained pixels and green additions in the edit panel. The edit adds 177,748 pixels with no removals, increasing IoU from 63.32% to 92.72%.

Modules 00–03 prepare the scene and region representations, module 04 illustrates fusion, and modules 05–09 describe OMR. The overview combines representation visualizations, fusion-weight diagnostics, and query predictions to explain their roles in the framework. Evaluation is performed after prediction. ScanNet reads point labels from class scores, while OVS uses language masks and object-informed spatial prediction.

Deletion and completion examples. For the main-paper "three cookies" example (teatime, frame 00140), OMR selects the L3 seed and accepts an alternate compact object candidate from L2. The final mask retains 9,149 seed pixels, removes 9,575 pixels, and adds none, increasing IoU from 49.34% to 94.08%. This illustrates how object-informed editing can remove excess extent around a semantically relevant target.

Figure S2 shows the complementary "sink" example (waldo_kitchen, frame 00140). The L1 seed misses part of the sink surface; OMR accepts an alternate compact object candidate from L3, retain-

ing 280,230 seed pixels and adding 177,748 pixels without removal. IoU increases from 63.32% to 92.72%. In both editing examples, the saved controller accepts an object-guided replacement and the final mask equals the accepted candidate; the added and removed regions are the set differences relative to the seed.

Visualization provenance. Across the three examples, representation renderings and evaluation-view fusion weight diagnostics illustrate their respective stages, while saved query responses and controller records supply the downstream masks and decisions. The overview combines newly rendered representation visualizations with saved downstream predictions to document the roles of the different stages. We independently verify the seed/final pixel-set identities and IoUs from the supplied masks. Reported IoUs and pixel counts use the full evaluation frame; display crops retain the complete union of the seed and final mask. Ground truth is used for evaluation and these diagnostic checks.

C Association and Weight Computation

Projection and visibility. Let T_v be the camera-to-world transform. The shared projection routine expects OpenGL axes and first flips its second and third rotation columns to OpenCV axes. With the resulting world-to-camera transform, compute $(x_{pv}, y_{pv}, z_{pv})^\top = [T_v^{-1}(\mathbf{x}_p, 1)^\top]_{1:3}$ and

$$(u_{pv}, v_{pv}) = \operatorname{round}\left(f_x x_{pv}/z_{pv} + c_x,\ f_y y_{pv}/z_{pv} + c_y\right). \tag{S6}$$

OVS checkpoint cameras already use OpenCV axes; their adapter supplies the corresponding OpenGL transform to this shared routine, avoiding a second axis conversion. Focal lengths are scaled to the cached region-map resolution. OVS uses the image center as principal point. Only finite points with $z_{pv} > 10^{-6}$ and in-image rounded coordinates are projected.

A 4×4 pixel bin defines a local depth reference $z_v^{\min}(u, v) = \min_{p':\mathrm{bin}(u_{p'v}, v_{p'v}) = \mathrm{bin}(u,v)} z_{p'v}$. The binding is accepted when the region ID is nonnegative and $\Delta z_{pv} = z_{pv} - z_v^{\min}(u_{pv}, v_{pv}) \le \epsilon_v$. We use $\epsilon_v = \delta_v$, sharing the admissibility tolerance and depth-decay scale. ScanNet uses $\delta_v = 0.05$ m. OVS uses

$$\delta_v = \max\left(10^{-6},\ 0.01 \operatorname{median}_w \|\mathbf{c}_w - \operatorname{median}_{w'} \mathbf{c}_{w'}\|_2\right), \qquad d_{pv} = \exp(-\Delta z_{pv}/\delta_v). \tag{S7}$$

The coordinatewise median defines the camera center used in this scale estimate. The local point z-buffer approximates visibility for association; alpha-composited feature rendering remains a separate operation.

Boundary agreement and instance counts. For $B_0 = 4$, define $\mathcal{D} = \{(\pm k, 0), (0, \pm k) : k = 1, \ldots, B_0\}$. With region map R_v and projected pixel $\mathbf{a}_{pv}$,

$$b_{pv} = 0.1 + \frac{0.9}{4B_0} \sum_{\mathbf{d} \in \mathcal{D}} \mathbb{1}[\mathbf{a}_{pv} + \mathbf{d} \text{ is in the image and } R_v(\mathbf{a}_{pv} + \mathbf{d}) = R_v(\mathbf{a}_{pv})]. \tag{S8}$$

The denominator stays 16 at image boundaries. Let $\mathcal{V}_v$ contain accepted bindings in view v, I_p be an estimated instance ID, and $r_{pv} = R_v(\mathbf{a}_{pv})$. Counts and compatibility are

$$\begin{aligned} n_{ivr} &= \sum_{p \in \mathcal{V}_v} \mathbb{1}[I_p = i,\ r_{pv} = r], \quad n_{iv} = \sum_r n_{ivr}, \\ h_{pv} &= \begin{cases} 0.25 + 0.75\, n_{I_p v r_{pv}}/n_{I_p v}, & I_p \ge 0, \\ 1, & I_p < 0. \end{cases} \end{aligned} \tag{S9}$$

Thus compatibility measures the visible fraction of an instance assigned to a region, not the fraction of a region belonging exclusively to that instance. Unknown instances are neutral. A large region can cover multiple instances and receive high compatibility for each; there is no dominant-instance rejection. Each carrier contributes at most once per view to these counts.

Semantic Anchoring and controls. For region feature $\mathbf{g}_{vr}$, $w_{pv} = d_{pv} b_{pv} h_{pv}$ and the finite baseline indicator m_p, the update is

$$\mathbf{e}'_p = \frac{\lambda m_p \mathbf{e}_p + \sum_{v: p \in \mathcal{V}_v} w_{pv} \mathbf{g}_{v r_{pv}}}{\lambda m_p + \sum_{v: p \in \mathcal{V}_v} w_{pv}}, \qquad \lambda = 1. \tag{S10}$$

An invalid baseline contributes zero numerator and zero prior mass. An unobserved row is copied exactly from the baseline, including an unknown row; a zero-total-mass row stays unknown. OVS features are normalized after rendering for query matching. The uniform, geometry, and fusion controls use respectively 1, db, and dbh on the same accepted bindings. No control changes the visibility or valid-point domain. With a finite baseline, the effective anchor coefficient is $\lambda/(\lambda + \sum_v w_{pv})$. Fixing λ across these controls therefore fixes initial mass, but does not fix the normalized anchor coefficient.

Instance construction and alignment. The matched OVS fields apply MiniBatchKMeans with 128 clusters to aligned, ℓ_2-normalized instance embeddings. The fitted partition assigns an instance label to every carrier. Instance coordinates and ordering are checked against the language Gaussians before association.

ScanNet uses the cached ViT-B-16 region/text space of `laion2B-s34B-b88K` and all 1,158 recorded views. Its cached masks have one effective semantic layer; repeated stored layers are not counted as independent evidence. Estimated instance labels are aligned by length and canonical point order; this alignment does not include an independent coordinate-based identity check.

Scope of the fusion controls. The matched controls evaluate the complete anchored update: depth and local-region weighting are applied jointly, and changing total observation weight also changes the effective anchor coefficient in Eq. 4. Their semantic-label metrics quantify the resulting predictions rather than the number of corrected geometric associations. The weights describe observation contributions, rather than calibrated uncertainty estimates.

D Complete Object-Guided Mask Refinement

D.1 Inputs, Candidates, and Cue-Informed Scores

The complete OMR consumes per-level response maps, masks, language confidences, and frozen native object assets. A language-confidence argmax provides the initial level b. Its eligible levels are L2/L3 on native 3D-OVS and L1/L2/L3 on native LERF-OVS; the object-informed selector can additionally rescue L1. The native mask thresholds are respectively 0.4 and 0.25. These thresholds define the dataset-specific native input masks.

Parameter conventions. We write scoring weights, acceptance thresholds, and comparison margins symbolically to describe their roles. They are empirically selected hyperparameters, fixed within each corresponding controlled comparison; they are not learned during prediction. Controller scores combine the specified cues and are not calibrated uncertainty estimates. Subscripts identify the cue or operation, and superscripts distinguish route-specific settings. An anonymized review version of our implementation code, training and evaluation scripts, and parameter configurations has been provided through an anonymous GitHub repository.

Native 3D object candidates. COS3D features yield per-Gaussian query relevance using the hard negatives and normalization in Appendix A. Greedy grouping repeatedly picks the highest-relevance unprocessed Gaussian whose relevance and opacity exceed τ_r and τ_o. Its normalized instance embedding is compared with all unprocessed embeddings using an ordered sequence of cosine thresholds $1 - \xi_t \bar{d}$, where $\bar{d}$ is the stored mean instance-feature distance. The first group whose opacity-weighted mean query relevance reaches τ_r is accepted. The last tested group is marked processed even if rejected; an empty group marks only its seed processed. Accepted groups are united, and absence of eligible seeds gives an empty proposal. Selected Gaussians are rendered with the native COS3D camera and rasterizer; thresholding rendered alpha at τ_α and nearest-neighbor resizing produce the candidate mask.

For a candidate C_ℓ, language mask M_ℓ, and response H_ℓ, define the language-relevance, mask-overlap, and spatial-alignment scores as

$$L_\ell = \mathrm{clip}_{[0,1]} \frac{\mathrm{mean}_{C_\ell} H_\ell}{\max H_\ell}, \qquad S_\ell = \frac{|C_\ell \cap M_\ell|}{\max(|M_\ell|, 1)}, \tag{S11}$$

$$P_\ell = \mathrm{clip}_{[0,1]} \left(w_b \frac{|C_\ell \cap \mathrm{band}(M_\ell)|}{\max(|M_\ell|, 1)} + w_d e^{-\kappa_d d_\ell} \right). \tag{S12}$$

The band is $\mathrm{dilate}_{k_d}(M) \setminus \mathrm{erode}_{k_e}(M)$, using elliptical kernels with fixed sizes k_d, k_e. The centroid distance d_ℓ is normalized by the image diagonal. Weights w_b, w_d balance boundary overlap and centroid proximity, and κ_d controls distance decay. Empty candidates or nonpositive response maxima give a zero language-relevance score. The centroid of an empty mask is $(0, 0)$; an empty 3D candidate has $G_\ell = 0$. Positive denominator floors are understood where needed for numerical stability. For selected 3D points with mean radius r, scene bounding-box diagonal D, and mean opacity $\bar{o}$, the geometry-informed and combined scores are

$$\begin{aligned} G_\ell = \mathrm{clip}_{[0,1]} \Big(& w_r e^{-\kappa_g r / \max(D, \epsilon)} \\ & + w_o \, \mathrm{clip}_{[0,1]} \frac{\bar{o} - o_0}{s_o} \Big), \\ U_\ell = w_L L_\ell + w_S S_\ell + w_G G_\ell + w_P P_\ell. & \end{aligned} \tag{S13}$$

Here w_r, w_o balance compactness and opacity; κ_g controls radius decay; o_0 is the opacity offset and $s_o > 0$ its scale; $\epsilon > 0$ stabilizes the denominator. The nonnegative weights w_L, w_S, w_G, w_P combine the language-relevance, mask-overlap, geometry-informed, and spatial-alignment scores. The weight sets (w_b, w_d), (w_r, w_o), and (w_L, w_S, w_G, w_P) each sum to one. The area ratio is $a_\ell = |C_\ell| / \max(|M_\ell|, 1)$.

D.2 Ordered Native Level Selection

Write c_ℓ for language confidence. Thresholds τ specify acceptance cutoffs, while δ and ε specify required gains and allowed deficits. The following tests are executed in order; the first accepted route wins.

1. **L1 rescue:** if $b \neq 1$, choose L1 when $c_1 - c_b \geq \delta_c^{(1)}$, $P_1 - P_b \geq \delta_P^{(1)}$, $L_1 \geq \tau_L^{(1)}$, and $U_1 \geq U_b - \varepsilon_U^{(1)}$.
2. **Background rescue:** the base is suspicious if $a_b \geq \tau_a^{(b)}$, $S_b \leq \tau_S^{(b)}$, or $U_b \leq \tau_U^{(b)}$. An alternate qualifies when $a_\ell \leq \tau_a^{(a)}$, $L_\ell \geq L_b$, $U_\ell - U_b \geq \delta_U^{(b)}$, and $S_\ell - S_b \geq \delta_S^{(b)}$. Among qualifying levels, maximize lexicographically $(U_\ell, S_\ell, -|a_\ell - 1|, L_\ell, c_\ell)$.
3. **L2/L3 rebalance:** for $b \in \{2, 3\}$ and the other level t, choose t when $\max(a_b, a_t) \leq \tau_a^{(r)}$, $U_t - U_b \geq \delta_U^{(r)}$, $S_t - S_b \geq \delta_S^{(r)}$, and $|a_b - 1| - |a_t - 1| \geq \delta_a^{(r)}$.
4. **Ambiguous L2/L3 switch:** choose t when $|c_t - c_b| \leq \varepsilon_c^{(a)}$ and $U_t - U_b \geq \delta_U^{(a)}$.
5. Otherwise retain b.

Stable iteration order resolves exact ties. The same selector is retained throughout each matched complete-OMR comparison. The output seed B is exactly the selected language mask.

D.3 Complete 3D-OVS Editing

Additional projected proposals. For the selected COS3D Gaussian set, let μ, r, and $\bar{f}$ be its centroid, mean radius, and normalized mean instance embedding. Each expansion retains the selected set and admits a point when its query score, opacity, and instance cosine with $\bar{f}$ pass fixed thresholds, and its distance to μ is at most $\max(sr, \rho D)$. The radius factor s and scene-scale floor ρ vary between the regular, coverage, and compact proposal profiles. An empty set stays empty. Expanded sets use the same native renderer.

Execution order and operation types. Initialize $Y = B$. Configured background, scene-object, same-level, and small-object routes are tested before the general coverage, compact-editing, and local-patching routes. Each branch reads the same frozen proposal bank without refitting features.

Coverage unites the language-level masks, checks the union-to-seed area ratio, and dilates the union when confidence and object evidence indicate incomplete coverage. The projected coverage proposal controls dilation size through its area ratio; its pixels are not directly united with the seed in this route.

Compact editing applies to an L1-rescue seed when geometric criteria and cross-level agreement satisfy the acceptance gates. It intersects the seed with a dilated union of the two-of-three level consensus and compact projection, then keeps the largest 8-connected component. Retained area and the responses of retained and removed pixels determine acceptance. This route contracts the mask.

Local patching filters both candidate and seed components using overlap, response, and size conditions. Candidate components that pass these tests can be added; seed components that fail them can be removed, with a response-based rescue for small components. The candidate filter is reapplied against the retained seed before their union. Other accepted routes can replace the whole mask. If no route is accepted, the seed is retained before final morphology.

The selected route determines whether a final binary dilation with an elliptical kernel is applied. Its size depends on mask extent and object-informed criteria; background-preservation routes bypass this operation. These morphology settings remain fixed within each component comparison.

D.4 Complete LERF-OVS Editing

Proposal construction and crop features. Each evaluation unit has a base-language proposal, an object-informed proposal, and, when available, a nonempty alternate compact-level proposal. The alternate maximizes a fixed weighted score combining normalized language confidence, bounding-box fill, largest-component fraction, area compatibility, and overlap with the base mask. Area compatibility is $\max(0, 1 - |a - a_{\mathrm{alt}}|/s_{\mathrm{alt}})$, where a is area relative to the base and $a_{\mathrm{alt}}, s_{\mathrm{alt}}$ set its preferred ratio and tolerance. Confidence is min–max normalized across levels; ties retain level order, and absence of an alternate falls back to the base proposal.

Proposals are aligned with the per-level 3D Gaussian sets selected for the query and grouped by query, source level, and Gaussian membership. The object feature is the normalized mean 3D language feature. Object confidence C_j is a weighted combination of geometric compactness, mean opacity, mask fill, largest-component fraction, and cross-view area consistency. Compactness decreases with the object-to-scene bounding-box diagonal ratio; area consistency decreases with the coefficient of variation of corresponding mask areas. Mask statistics are averaged across contributing views.

For each object, up to a fixed maximum number of distinct views is selected by area ratio, fill, and largest-component fraction. Expanded proposal crops have their exteriors whitened and are resized for the frozen CLIP encoder. Normalized crop vectors are pooled with area-ratio weights and the result is normalized; pairwise cosine similarity measures view consistency. Candidate construction and crop pooling are held fixed throughout each matched comparison.

Scores and replacement gates. For candidate j, query relevance p_j, query cosine s_j, confidence C_j, view consistency V_j, image area ratio A_j, and fill F_j give

$$J_j = \boldsymbol{\theta}^\top \boldsymbol{\phi}_j, \qquad \boldsymbol{\phi}_j = [p_j, s_j, C_j, V_j, A_j, F_j]^\top. \tag{S14}$$

This specifies the crop-based proposal ranking summarized in Section 3.3. For local patch acceptance, OMR assigns the selected seed mask B a *seed-mask reference score* $Q = \nu_c \bar{c} + \nu_F F_B + \nu_R R_B + \nu_n / \max(n_B, 1)$, where F_B is the seed's bounding-box fill ratio, R_B is the fraction of seed pixels in its largest 8-connected component, n_B counts its 8-connected components, and $\bar{c}$ is the selected level's language confidence after min–max normalization across levels. This score provides a reference for evaluating a candidate edit to the seed. A degenerate confidence range assigns a shared neutral value to all levels. The fixed nonnegative vectors $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ each sum to one. A nontrivial proposal is preferred when its J is within ε_J of the best candidate; otherwise the maximum is retained. Ties compare relevance, cosine, confidence, and area ratio.

The replacement score is

$$\begin{aligned} T_j = \lambda_n \Delta_n^+ + \lambda_R \Delta_R^+ + \lambda_F \Delta_F^+ \\ + \lambda_a \max(0, 1 - |a_j - a_\star|/s_a) \\ + \lambda_p p_j + \lambda_C C_j + \lambda_q \mathbb{1}[\text{bestquery}(j) = q]. \end{aligned} \tag{S15}$$

Here $a_j = |O_j|/\max(|B|, 1)$ for candidate mask O_j, $\Delta_n^+ = (n_B - n_j)_+/\max(n_B, 1)$, and Δ_R^+, Δ_F^+ are positive gains in largest-component fraction and fill. The preferred area ratio $a_\star$ and positive tolerance s_a define area compatibility. The nonnegative replacement weights $\lambda_n, \lambda_R, \lambda_F, \lambda_a, \lambda_p, \lambda_C, \lambda_q$ sum to one; they are separate from the fusion baseline mass λ.

Regular replacement requires an alternate compact proposal whose best query is q, $p_j \geq \tau_p^{\text{rep}}$, and $n_j \leq n_B$. The seed must fail the strong-large-object test and exhibit insufficient fill, extent, connectivity, or excessive fragmentation. IoU and area ratio must lie within fixed acceptance intervals, and $T_j \geq \tau_T^{\text{rep}}$. These gates prevent a high score alone from triggering replacement. Realignment, carve-out, and small-object routes use their corresponding structural acceptance gates.

Ordered editing and fallbacks.

1. Test every object for replacement. Among accepted objects, maximize (T_j, p_j, C_j) lexicographically and replace the whole seed.
2. If none passes, test the preferred object for a local patch. Acceptance requires sufficient overlap and language relevance, a bounded area change, nonincreasing fragmentation, and controlled losses of connectivity and fill. Candidate score must also exceed the seed-mask reference score Q by the required margin. Overwrite the seed within the object bounding box; this may add or delete pixels. A pixel-identical patch leaves the route inactive.
3. If still inactive, test compact score-map proposals and their structural gates, followed by component proposals and the fixed best-mix fallback. The first accepted fallback replaces the mask.
4. If all tests fail, return B without editing.

Score-map proposals use fixed response thresholds and 8-connectivity. The compact builder keeps the largest component and restricts both image-area fraction and proposal-to-seed area ratio. A weighted score combines mean and peak response, fill, fill improvement, and area compatibility. Acceptance requires a dominant connected component and sufficient mean and peak response.

Fixed prediction settings. Candidate construction, scoring coefficients, selection margins, editing gates, and fallback order are held fixed within each component comparison. Only the specified prediction-stage switch changes.

D.5 Component Ablations of the Native Complete Framework

Main-text Table 6 uses the native inputs of the final 3D-OVS and LERF-OVS systems. Let S denote the recorded native object-informed selector and E the complete editing stage. For $S = 0$, the seed is the native language-confidence choice (L2/L3 on 3D-OVS, L1/L2/L3 on LERF), and the routing state becomes seed retention; all subsequent editing decisions are recomputed. For $E = 0$, the selected seed is passed through without executing the editing stage. This bypass removes completion, replacement, contraction, morphology, and fallback operations together; it does not isolate their individual internal rules. In particular, $E = 0$ retains native object-informed selection when $S = 1$. This differs from the *Fusion only* configuration in Table 5, which disables both OMR selection and editing. The distinction is the component switch: removing editing alone retains selection, whereas Fusion only disables OMR entirely.

The selector supplies a seed for subsequent editing, so its contribution is assessed with the editor enabled. The selected seed need not improve standalone mask accuracy. On LERF-OVS, selection without editing gives 54.7334% mIoU compared with 59.1284% for Fusion only, whereas selection with editing improves the final output from 67.9195% to 68.4450%. These results distinguish seed accuracy from the effect of selection on the complete OMR output.

Table S2: Per-scene final-system prediction ablations corresponding to main-text Table 6 (mIoU, %). Each removal retains the other stages. Bold indicates the highest value in each row, including ties.

Dataset	Scene	Full	w/o S	w/o E
3D-OVS	bed	**93.1743**	86.9163	88.3158
3D-OVS	bench	**95.3129**	90.0380	95.1666
3D-OVS	lawn	**97.3005**	97.2461	97.1743
3D-OVS	room	**95.7365**	**95.7365**	86.9680
3D-OVS	sofa	**92.8640**	76.8102	90.6053
LERF-OVS	figurines	**63.8688**	63.4018	54.9128
LERF-OVS	ramen	**68.0161**	64.8581	49.9676
LERF-OVS	teatime	**77.9845**	77.9303	64.6881
LERF-OVS	waldo_kitchen	63.9105	**65.4880**	49.3651

E Staged Training and Optimization Scope

Hardware and fixed settings. Representation training uses one NVIDIA L40 GPU; cached fusion controls and diagnostics run on CPU. RGB reconstruction and COS3D instance learning each use 30,000 iterations; language-field training uses 10,000 iterations, three levels, 64 dictionary vectors, and top-$k = 4$ decoding. Fusion uses $\lambda = 1$, $\eta = 0.25$, $\beta = 0.1$, and a cross-shaped local-region neighborhood of radius four pixels. Refinement freezes geometry and representation-model parameters, and controls reuse the same trained representations.

Inherited representation training. We construct the upstream representations with the standard LangSplatV2 and COS3D training recipes (Li et al., 2025; Zhu et al., 2025). Image preprocessing uses SAM masks and CLIP region embeddings at three semantic levels. RGB Gaussian reconstruction is trained first; language and instance training then initialize from that reconstruction. LangSplatV2 optimizes the sparse language coefficients and codebook vectors, and COS3D optimizes the per-Gaussian instance features. Both feature stages keep Gaussian positions, scale, rotation, opacity, and RGB appearance fixed. The SAM and CLIP encoders remain frozen.

Table S3: Default inherited representation-training recipe. The iteration budget is per scene and, for feature training, per semantic level.

Stage	Iterations	Optimized parameters
RGB reconstruction	30,000	Geometry, opacity, RGB appearance
LangSplatV2 language	10,000	Sparse logits and codebook vectors
COS3D instance	30,000	Instance features
Fusion and OMR	0	No gradient-based optimization

Preprocessing and initialization. Preprocessing uses pretrained SAM ViT-H and frozen 512-dimensional OpenCLIP ViT-B-16 descriptors from `laion2B-s34B-b88K`. Camera poses and intrinsics come from scene reconstruction; per-level region maps and descriptors are retained. Feature stages start from the 30,000-iteration RGB reconstruction. Language training runs for 10,000 iterations using 64 dictionary vectors, top-$k = 4$ decoding, and levels L1/L2/L3. The instance–language mapping uses fixed local neighbor association and spatial smoothing on the stored features. Optimizer and reconstruction schedules follow the upstream methods (Kerbl et al., 2023; Li et al., 2025; Zhu et al., 2025); densification is disabled during feature-only training. No additional encoder-training stage is introduced.

Integration and fixed scope. The fusion stage aggregates the trained language features using weights informed by geometry and instance structure, and OMR reads the resulting fields without retraining the encoders, Gaussian geometry, language codebooks, or instance features. The component ablations reuse the same representation checkpoints in every row; only the specified fusion/OMR switches change. Each comparison uses the stage-specific training budgets in Table S3.

The training budget describes representation construction; component ablations vary the integration and prediction stages without retraining.

F Datasets and Evaluation Protocols

F.1 Rendered Query Masks

The 3D-OVS evaluation contains bed, bench, lawn, room, and sofa, with 30/35/30/30/30 query–frame masks respectively. The LERF-OVS dense-view protocol contains figurines, ramen, teatime, and waldo_kitchen, with 56/71/59/22 masks. A query–frame pair is one evaluation unit. For scene s with evaluated units $\mathcal{U}_s$, we compute

$$\mathrm{mIoU}_s = \frac{1}{|\mathcal{U}_s|} \sum_{u \in \mathcal{U}_s} \frac{|Y_u \cap G_u|}{|Y_u \cup G_u|}. \tag{S16}$$

Following COS3D's LERF-OVS segmentation metric definitions (Zhu et al., 2025), we also compute

$$\mathrm{mAcc}_s = \frac{1}{|\mathcal{U}_s|} \sum_{u \in \mathcal{U}_s} \mathbb{1}[\mathrm{IoU}(Y_u, G_u) > 0.25]. \tag{S17}$$

Dataset results are unweighted means of scene results. They are not pooled means over the 155 or 208 masks. Percentages multiply these values by 100.

Scoring convention and 3D consistency. Our LERF-OVS mIoU, mAcc, and scene averaging follow COS3D. Its prediction procedure selects 3D Gaussians and renders the shared selection across views, imposing a stronger 3D consistency constraint than independent viewwise predictions. This motivates our use of its segmentation scoring convention. Our method uses 3D instance structure to inform query–frame mask refinement; sharing the metrics does not imply identical prediction generation. LangSplatV2's separately reported localization accuracy uses a different criterion, as detailed in Appendix H.3.

Mask alignment and evaluated queries. PNG masks are converted to grayscale and thresholded at 127. Ground truth is resized to the prediction resolution by nearest-neighbor interpolation: OpenCV for 3D-OVS and Pillow for all LERF component configurations and the complete framework. The additional LangSplatV2-aligned scoring comparison uses OpenCV to rasterize the original LERF polygon annotations (Appendix H.3). A zero union contributes zero IoU in these main evaluation routines. No per-query winner is selected across levels or variants during scoring. Frame IDs and query strings are matched exactly; queries absent from a frame's annotation list are not scored there. The evaluation comprises 363 valid frame–query pairs.

F.2 Native-Point ScanNet Evaluation

Following the experimental protocol of OpenGaussian, we evaluate the same ten selected ScanNetV2 scenes: scene0000_00, scene0062_00, scene0070_00, scene0097_00, scene0140_00, scene0200_00, scene0347_00, scene0400_00, scene0590_00, and scene0645_00. Together they provide 1,158 camera views and 1,569,975 native points. All valid GT points belonging to a given taxonomy remain in the evaluation domain. Neither opacity nor observation coverage filters remove them. A nonfinite score row remains unknown when no valid observations are available.

For the classes $\mathcal{C}_s$ present in a scene's ground truth, the metrics are

$$\mathrm{mIoU}_s = \frac{1}{|\mathcal{C}_s|} \sum_{c \in \mathcal{C}_s} \frac{\mathrm{TP}_{s,c}}{\mathrm{TP}_{s,c} + \mathrm{FP}_{s,c} + \mathrm{FN}_{s,c}}, \tag{S18}$$

$$\mathrm{mAcc}_s = \frac{1}{|\mathcal{C}_s|} \sum_{c \in \mathcal{C}_s} \frac{\mathrm{TP}_{s,c}}{\mathrm{TP}_{s,c} + \mathrm{FN}_{s,c}}, \qquad \mathrm{accuracy}_s = \frac{\#\text{correct valid points}}{\#\text{valid GT points}}. \tag{S19}$$

Dataset values average the scene metrics equally. Unknown predictions remain errors on their valid GT points. The 10-class taxonomy contains wall, floor, bed, chair, sofa, table, door, window, bookshelf, and toilet. The 15-class protocol additionally includes cabinet, counter, desk, curtain, and sink. The 19-class protocol additionally includes picture, refrigerator, shower curtain, and bathtub. Text-score columns follow the corresponding recorded NYU40 class IDs before argmax.

Label mapping and averaging order. The ordered NYU40 IDs are $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 24, 28, 33, 34, 36]$ for 19 classes, $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 33, 34]$ for 15 classes, and $[1, 2, 4, 5, 6, 7, 8, 9, 10, 33]$ for 10 classes. For each taxonomy, valid GT labels are mapped to its ordered score columns; other GT labels are ignored. Predicting an unknown label on a valid GT point is an error, not a removal of that point. Metrics first count pointwise TP, FP, and FN, then average over GT-present classes within each scene, and finally over the ten scenes. Thus neither class weighting across the entire dataset nor pooling all native points reproduces the reported mean.

External baseline sources and evaluation domains. The baseline values in the main comparisons are taken from external reports. For ScanNet, the upper comparison block uses OpenGaussian's scene and category sets. OpenGaussian and OpenInsGaussian discard low-opacity points in their released evaluators, whereas our evaluation retains all valid ground-truth points. Dr. Splat and EmbodiedSplat use other settings and are listed separately; the EmbodiedSplat row is its RGB-input SAM+CLIP variant. The reported values thus retain their respective evaluation domains. The matched fusion controls use the same native-point domain and readout for all compared configurations.

G Optimized ScanNet Probability Readout

Requested-taxonomy probabilities before fusion. For taxonomy $\mathcal{C}$, let s^0_{ic} and s^{vr}_c be cosine similarities of the native baseline and a region descriptor to the v1 text bank. Normalize each source within this taxonomy before accumulation:

$$p^0_{ic} = \frac{\exp(s^0_{ic}/T)}{\sum_{d\in\mathcal{C}} \exp(s^0_{id}/T)}, \quad p^{vr}_c = \frac{\exp(s^{vr}_c/T)}{\sum_{d\in\mathcal{C}} \exp(s^{vr}_d/T)}, \qquad T = 0.03. \tag{S20}$$

Apply the same fusion bindings and $w_{iv} = d_{iv} b_{iv} h_{iv}$ as in Appendix C:

$$P_{ic} = \frac{p^0_{ic} + \sum_v w_{iv} p^{vr_{iv}}_c}{1 + \sum_v w_{iv}}. \tag{S21}$$

Invalid baselines and unobserved rows follow Appendix C. Each taxonomy is normalized independently before fusion; columns from the 19-, 15-, and 10-class spaces do not compete with one another.

One spatial averaging step. For each native point, take its 16 nearest neighbors, including itself, and define

$$a_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2(0.06)^2} - \frac{\|\mathbf{I}_i - \mathbf{I}_j\|_2^2}{2(40)^2}\right), \qquad \bar{P}_{ic} = \frac{\sum_{j\in\mathcal{N}_i} a_{ij} P_{jc}}{\sum_{j\in\mathcal{N}_i} a_{ij}}. \tag{S22}$$

Coordinates are in meters and colors use their native image scale. For an observed point with valid neighboring predictions, set $\widetilde{P}_i = 0.5P_i + 0.5\bar{P}_i$; otherwise retain P_i. Only finite neighbor scores contribute to the average.

Unlabeled priors, calibration, and final labels. Estimate priors from the unsmoothed fused probabilities:

$$\pi_c = \max\left(\epsilon_\pi, \frac{1}{10}\sum_s \frac{1}{|\mathcal{F}_s|}\sum_{i\in\mathcal{F}_s} P_{ic}\right), \quad Q_{ic} = \frac{\widetilde{P}_{ic}}{\pi_c^{0.25}}, \qquad \widehat{y}_i = \arg\max_{c\in\mathcal{C}} Q_{ic}. \tag{S23}$$

Here $\epsilon_\pi > 0$ is a numerical floor, and $\mathcal{F}_s$ contains finite native rows independent of GT labels, opacity, and observation coverage. Priors therefore average scenes equally. Prior calibration follows spatial averaging and can change finite unobserved points; final labels are obtained by taxonomy-wise argmax.

H MATCHED FRAMEWORK ABLATIONS AND ADDITIONAL RESULTS

This appendix gives mass-matched fusion controls, observation diagnostics, per-scene point-prediction results, and annotation-scoring comparisons. The final native-system prediction ablation is detailed in Appendix D.5. The comparisons are grouped by their fixed inputs and prediction settings so that their relation to the main-paper results is explicit.

H.1 MASS-MATCHED FUSION CONTROLS AND OBSERVATION DIAGNOSTICS

Predictions are frozen before GT is used for scoring or diagnostics; all valid native GT points are retained, including those without accepted observations. Full improves over mass-matched Uniform in 8, 9, and 7 of ten scenes for the 19-, 15-, and 10-class protocols, respectively. The category-correct observation share rises in nine of ten scenes. The improvement example below includes 175 incorrect-to-correct and 33 correct-to-incorrect changes from Uniform to Full.

The controls use ten ScanNet scenes, 1,158 views, and 1,569,975 native points. All accepted associations, features, hard geometric filtering, and cosine readout are shared, with $\lambda = 1$. The uniform mass-matched weight is W_p/K_p per accepted observation, retaining Full's per-point observation mass and effective anchor proportion for finite initial vectors. No-observation, invalid-anchor, and zero-mass cases follow the original fallback rules. Stored observation masses match exactly; independent coefficient reconstruction has maximum relative error 5.78×10^{-8}.

Table S4: Fusion controls on ScanNet (%). The first block isolates relative observation weighting at equal per-point observation mass and effective anchor proportion. The second gives reference and factor-control results. All fusion configurations share hard geometric admissibility and cosine readout. Bold marks the highest value within each block, not across blocks. d: depth consistency; b: local region agreement; h: instance compatibility.

Configuration	19 mIoU	19 mAcc	15 mIoU	10 mIoU
Matched-mass comparison: relative weighting				
Uniform / matched mass	36.7398	56.4979	36.0969	44.3463
Full (dbh)	**37.6899**	**57.6239**	**37.0852**	**45.6601**
Additional reference and factor controls				
Original	33.4675	50.6005	32.8591	41.1075
Uniform	36.7840	56.8322	36.1381	44.4393
db (without h)	36.8816	56.8764	36.2330	44.6027
bh (without d)	37.6605	57.6234	37.0490	45.5998
dh (without b)	**37.7226**	57.6635	**37.1126**	**45.7009**
h only	37.6930	**57.6670**	37.0790	45.6388

The primary comparison is the matched-mass block in Table S4. Full exceeds mass-matched Uniform in all four reported metrics, with mIoU gains of 0.9501, 0.9883, and 1.3139 points for the three taxonomies. This supports the benefit of relative observation weighting beyond changes in the effective anchor proportion. The additional controls characterize the weighting factors: h only is close to Full, and dh has slightly higher mIoU than Full in all three taxonomies. Factor removal changes observation mass as well as allocation, so these rows do not isolate fixed-anchor effects of individual factors or establish necessity of all three soft factors.

For diagnostics, a teacher label is the unfused region cosine argmax over the same 19 classes. We compare the unweighted share of correct teacher observations with their share of total fusion weight, using valid GT points with finite observations and positive total mass. Point means are computed within each scene, followed by equal-scene averaging. The shares are 45.2868% and 46.1076%, respectively; coverage averages 96.1384% of valid GT points. The share omits the anchor and continuous score magnitudes, so it does not determine the final prediction. Independent recomputation of 170 sampled points has maximum share error 1.11×10^{-16}.

The improvement example is scene0645_00, frame 1420, region 11, with 4,227 valid points. Candidates require at least 1,000 points, largest GT class share at most 0.8, Full accuracy above Original and Uniform, and at least 20 Uniform-to-Full corrections. The lower median of 968 qualifying cases is selected by Full-minus-Uniform gain, with scene/frame/region tie-breaking. The rule was written

after aggregate fusion results were known; this is a post-hoc explanatory example. Its correct-observation share rises from 33.9783% to 37.1689%. The fixed failure case, scene0140_00, frame 2520, region 18, has 25,611 points and a share decrease from 35.6251% to 35.3018%.

Table S5: Accuracy on the same native points within each diagnostic region (%); these values are not scene mIoU or image-pixel accuracy.

Case	Original	Uniform	Full	Matched Uniform
Improvement	41.5188	72.1552	75.5145	71.8476
Failure	45.9451	50.1464	45.7811	51.8059

The failure includes 951 incorrect-to-correct and 2,069 correct-to-incorrect changes from Uniform to Full. These cases document both outcomes under frozen predictions; they do not isolate descriptor, association, and propagation errors from one another.

H.2 Per-Scene Point Prediction and Fusion Diagnostics

Table S6: Per-scene breakdown of the GAFF point-prediction configuration's 19-class ScanNet result in main-text Table 3 (mIoU, %). Optimized includes fusion, probability readout, spatial averaging, and class-prior calibration; its scene mean is 38.2770%. Original has a scene mean of 33.4675%. Increments are calculated from the underlying values and displayed to four decimal places.

Scene	Original	Optimized	Δ (pp)
scene0000_00	26.5696	**39.2572**	+12.6876
scene0062_00	24.4764	**34.9894**	+10.5130
scene0070_00	33.4514	**36.8837**	+3.4323
scene0097_00	38.2249	**40.7620**	+2.5371
scene0140_00	**36.5034**	32.4005	−4.1029
scene0200_00	31.7782	**42.3719**	+10.5936
scene0347_00	42.4882	**42.9181**	+0.4299
scene0400_00	30.7580	**36.2458**	+5.4878
scene0590_00	39.0078	**42.6260**	+3.6182
scene0645_00	31.4167	**34.3157**	+2.8990

Fusion-only diagnostic corresponding to Table 4. The following analysis concerns the GAFF fusion stage (dbh) configuration with fixed cosine-score readout in main-text Table 4, whose 19-class mean is 37.6899%. It diagnoses that fusion configuration, not the optimized point-prediction results in the preceding table. On scene0140, door/picture/floor IoU decreases by 21.5283/16.8654/8.2456 pp. There are 40,786 correct-to-incorrect point changes and 40,955 changes in the opposite direction. Similar point counts do not imply preserved class-mean IoU. The mean per-point weighted fraction of correct teacher observations is 17.59% for newly incorrect points and 56.21% for corrected points. Newly incorrect points average 13.67 observations: repeated observations can reinforce a consistently incorrect teacher response. These use GT after prediction and are not available to the inference weights. In frame 2520, region 18 contains 25,611 valid points; its largest GT class, floor, covers only 48.11%, while the teacher predicts table. This case shows the co-occurrence of mixed region coverage and incorrect teacher semantics, motivating further improvements in observation reliability.

H.3 LERF-OVS Segmentation Scoring with LangSplatV2 Annotations

To compare segmentation quality using the annotation handling in LangSplatV2's official evaluation, we re-score the same 208 saved final masks against the original polygon ground truth rasterized with OpenCV. The query–frame pairs and predictions are unchanged. We average query–frame IoUs within each scene and then average the four scenes equally. Table S7 compares these results with the LangSplatV2 values reported in Table 2 of its original paper (Li et al., 2025); we do not rerun the LangSplatV2 baseline.

Table S7: LERF-OVS segmentation mIoU (%) with LangSplatV2's polygon-ground-truth rasterization. LangSplatV2 values, including its rounded mean, are quoted from the original paper. Our predictions are fixed and re-scored. Scene gains use the displayed values; the mean gain is relative to the published 59.9% mean.

Scene	LangSplatV2 (reported)	GAFF (re-scored)	Δ (pp)
figurines	56.4	**63.8640**	+7.4640
ramen	51.8	**68.0628**	+16.2628
teatime	72.2	**78.0305**	+5.8305
waldo_kitchen	59.1	**64.0076**	+4.9076
Scene mean	59.9	**68.4912**	$\approx +8.59$

The GAFF framework exceeds the reported LangSplatV2 segmentation mIoU in all four scenes, with an approximately 8.59 pp higher scene mean. The change from 68.4450% in the main evaluation to 68.4912% here is due to boundary-pixel differences between Pillow and OpenCV polygon rasterization, rather than different predictions. The published LangSplatV2 mean is retained at its original precision; it is not recomputed from the rounded per-scene values.

Segmentation and localization are distinct. This comparison evaluates final segmentation masks. LangSplatV2's localization procedure smooths continuous relevance maps, selects their peak and semantic level, and checks whether the predicted point lies inside an annotated bounding box (Li et al., 2025). Its reported 84.1% localization accuracy is therefore not directly comparable to our 90.4426% segmentation mAcc. We evaluate segmentation using the saved binary masks; localization instead requires continuous relevance maps and its own point-in-box criterion. Accordingly, this comparison reports segmentation mIoU.