RELIASPLAT: RELIABLE FEATURES IN LANGUAGE GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary 3D scene understanding commonly builds language fields by distilling multi-view 2D language features into 3D representations. In Language Gaussian Splatting, these pre-extracted features serve as the basic semantic observations for constructing Gaussian language fields. However, their accuracy is often affected by occlusions, scale variations, region-partition errors, and noise from vision-language models. As a result, the same 3D entity may receive inconsistent or incorrect semantic observations across views. Existing methods mainly focus on how language features are represented or aggregated in 3D, while the reliability of the input features themselves receives less attention. ReliaSplat is a training-free framework for improving input language features. Cross-View Reliability-Aware Aggregation estimates observation reliability from cross-view semantic consistency and suppresses inconsistent features during aggregation. Consensus-Guided Feature Refinement uses the resulting 3D semantic consensus to correct noisy 2D features through a lightweight 3D→2D→3D process. Geometry-Aligned Semantic Propagation further supplements reliable semantic observations. Experiments on LERF and ScanNet show consistent improvements in open-vocabulary semantic segmentation, with an average mIoU of 64.3% on LERF. Cross-framework experiments further demonstrate that the enhanced language features can directly improve multiple Language Gaussian Splatting methods, validating the generality of the feature enhancement framework.

1 INTRODUCTION

Open-vocabulary 3D scene understanding aims to construct 3D semantic representations that can respond to arbitrary natural-language queries. With the development of vision-language models (Radford et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023), language features extracted from multiple images can be distilled into explicit 3D representations. LERF and LangSplat have demonstrated the effectiveness of this paradigm for open-vocabulary localization and segmentation (Kerr et al., 2023; Qin et al., 2024). Compared with closed-set semantic representations, language fields preserve rich vision-language embeddings and support flexible text queries without predefined categories. In Language Gaussian Splatting, pre-extracted 2D language features therefore serve as the basic semantic observations for constructing 3D language fields. However, the quality of these observations directly affects the resulting 3D semantics. Such features are commonly extracted from image regions, segmentation masks (Kirillov et al., 2023), or multi-scale crops. Their semantic content can be affected by occlusion, scale variation, region-partition errors, and noise from vision-language models. Consequently, the same physical entity may receive inconsistent or incorrect language features across different views.

Previous methods mainly focus on the representation, distillation, or aggregation of language features in 3D, while the reliability of the input semantic observations themselves receives less attention. Once an unreliable feature is accepted as valid semantic evidence, its error can be inherited by the corresponding Gaussian representation and further propagated to rendered semantic maps and downstream localization or segmentation. Improving the 3D representation alone cannot fully remove errors that already exist in the input language features. This issue is particularly important during multi-view aggregation. Previous methods often use visibility or rendering significance to measure observation importance. These quantities reflect geometric or rendering importance rather than semantic reliability. A high rendering significance does not guarantee semantic correctness. In

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.

contrast, semantic agreement among observations of the same 3D entity provides direct evidence of semantic stability. Features that remain consistent across views are more likely to describe stable semantics, while observations that strongly disagree with others are more likely to be affected by occlusion, region errors, or feature noise. Cross-view semantic consistency therefore provides a natural reliability cue without additional supervision or network training.

Based on this observation, we propose ReliaSplat, a training-free framework for enhancing input language features by explicitly modeling the reliability of semantic observations. Cross-View Reliability-Aware Aggregation estimates observation reliability from cross-view semantic consistency and combines it with rendering significance to suppress inconsistent observations during 3D aggregation. Consensus-Guided Feature Refinement uses the 3D semantic consensus constructed from reliable multi-view evidence to correct noisy 2D language features through a lightweight 3D→2D→3D process. The correction strength is adaptively determined by the agreement between each 2D observation and the 3D consensus. Stable semantics are therefore preserved, while inconsistent features receive stronger correction. On this basis, Geometry-Aligned Semantic Propagation reuses reliable language features from real views and supplements valid semantic observations under geometric and semantic consistency constraints.

ReliaSplat improves the input language features themselves, allowing the enhanced features to be used across different Language Gaussian Splatting frameworks. Experiments on LERF and ScanNet show consistent improvements in open-vocabulary semantic segmentation, with an average mIoU of **64.25%** on the four LERF scenes. Ablation studies further validate the effects of reliability-aware aggregation, consensus-guided refinement, and geometry-aligned propagation. Cross-framework experiments strictly follow the official pipelines, code, and evaluation protocols of LangSplat, LangSplatV2, VALA, and THGS, with only the input language features replaced. The refined features improve the mean mIoU across all four downstream frameworks while keeping the mean mAcc stable overall. These results indicate that ReliaSplat improves general input language-feature quality rather than a specific downstream aggregation rule.

The main contributions are summarised as follows.

- Cross-View Reliability-Aware Aggregation is proposed to estimate observation reliability from cross-view semantic consistency and combine it with rendering significance, thereby suppressing inconsistent observations during 3D aggregation.
- Consensus-Guided Feature Refinement is proposed to use reliable 3D semantic consensus to correct noisy 2D language features through a lightweight 3D→2D→3D process, improving feature quality at the observation level. Geometry-Aligned Semantic Propagation further reuses reliable real-view features under geometric and semantic consistency constraints to supplement semantic observations.
- ReliaSplat provides a training-free framework for improving input language features. The enhanced features can be directly integrated into multiple Language Gaussian Splatting methods, demonstrating cross-framework generalization.

2 Related Work

2.1 3D Language Fields

Open-vocabulary 3D scene understanding links natural-language semantics to 3D scenes. This supports open-vocabulary localization and segmentation. Unlike fixed category labels, language fields store continuous vision-language embeddings. The embeddings are attached to spatial locations or 3D primitives. Text queries are encoded in the same feature space. Their similarity to scene embeddings produces relevance responses. Depending on the representation, the field supports spatial queries or image-space response maps. Features from multiple posed images are mapped into a shared spatial representation. The resulting field describes scene-level semantics. It does not produce unrelated predictions for each image. LERF distills vision-language features into a NeRF representation (Mildenhall et al., 2020). This yields a continuous language field for 3D text queries (Kerr et al., 2023). LangSplat extends this approach to 3D Gaussian Splatting. It attaches multi-scale language features to Gaussian primitives and constructs an explicit language field (Qin et al., 2024). The two methods represent and render features differently. LERF uses volumetric rendering along camera rays. LangSplat renders Gaussian-associated features into image space.

LERF uses multi-scale CLIP supervision to support queries at different spatial scales. LangSplat uses segmentation-informed hierarchical semantics to distinguish object regions. Both methods associate multi-view 2D language features with the scene. The same text query can then be evaluated from different viewpoints.

2.2 Language Gaussian Splatting

Training-free Language Gaussian Splatting methods avoid additional semantic optimisation. They directly aggregate or propagate pre-extracted 2D language features. Occam's LGS constructs language Gaussians through direct feature aggregation (Cheng et al., 2025). VALA improves multi-view aggregation without scene-specific language training (Wang et al., 2026). ReLaGS also builds 3D language representations from existing semantic observations (Xie et al., 2026). THGS constructs a training-free superpoint graph for hierarchical scene understanding (Dai et al., 2025).

Training-based approaches address different parts of language-field learning. LangSplat learns scene-specific language representations by optimising semantic features associated with 3D Gaussians (Qin et al., 2024). 3D VL-GS combines cross-modal rasterization with camera-view blending to improve semantic rendering and consistency (Peng et al., 2025). GAGS uses camera-distance-aware segmentation. Its granularity factor selects consistent 2D features during distillation (Peng et al., 2026). LangSplatV2 accelerates high-dimensional feature rendering and text queries (Li et al., 2025). LaGa models view-dependent semantics through object-level cross-view associations (Cen et al., 2025). OpenGaussian connects 3D points with 2D semantic features for point-level understanding (Wu et al., 2024). Together, these approaches improve feature learning, semantic rendering, and cross-view association in language fields.

These methods differ in feature representation and aggregation. Most still treat pre-extracted 2D language features as fixed semantic observations. Occlusion, scale variation, region-partition errors, and vision-language model noise can corrupt those features. The same 3D entity may then receive inconsistent or incorrect semantics across views. Distillation or aggregation may carry these errors into the 3D language field. ReliaSplat addresses the reliability of input semantic observations. It uses cross-view semantic consensus to refine inconsistent features during field construction. Geometry-Aligned Semantic Propagation reuses reliable real-view features as complementary semantic evidence.

3 Problem Formulation

Language Gaussian Splatting constructs 3D language fields from language features extracted across multiple views. Let $\{G_i\}_{i=1}^{N_g}$ denote the 3D Gaussians. For Gaussian i observed in view v, its language feature is denoted by $f_{iv} \in \mathbb{R}^D$, with the normalized representation $\hat{f}_{iv} = f_{iv}/\|f_{iv}\|_2$. Each Gaussian is therefore associated with a set of multi-view semantic observations, $\mathcal{F}_i = \{\hat{f}_{iv}\}_{v\in\mathcal{V}_i}$.

Ideally, observations of the same 3D entity should preserve consistent semantics across views. In practice, occlusion, scale variation, region-partition errors, and vision-language model noise can introduce semantic deviations. Existing aggregation commonly relies on visibility or rendering significance to measure observation importance. Let s_{iv} denote the rendering significance of Gaussian i in view v, measured by its alpha-composited contribution at the projected center, and let r_{iv} denote its unobservable semantic reliability. These quantities describe different properties.

A large rendering significance does not guarantee semantic correctness. Unreliable observations may therefore receive large aggregation weights and propagate errors into the 3D language field. The objective is to identify reliable semantic evidence from existing multi-view observations and use it to enhance unreliable language features. This process requires no additional network training. The enhanced features can be directly applied to different Language Gaussian Splatting frameworks.

4 METHOD

4.1 OVERVIEW

ReliaSplat formulates language feature enhancement in Language Gaussian Splatting as a semantic observation reliability problem. The goal is to identify reliable semantics from multi-view observations and use cross-view consensus to refine inconsistent features. On this basis, reliable semantic observations are further extended.

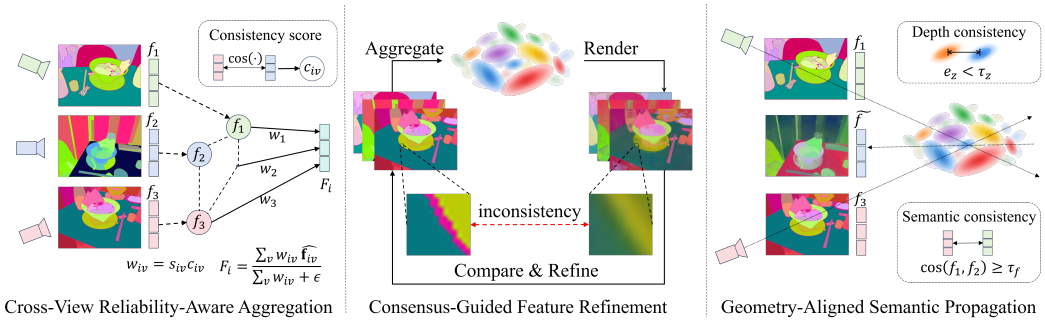


Figure 1: Method overview. Left: Cross-View Reliability-Aware Aggregation estimates observation reliability from cross-view consistency and constructs a 3D semantic consensus. Middle: Consensus-Guided Feature Refinement projects the 3D consensus back to 2D and corrects noisy features. Right: Geometry-Aligned Semantic Propagation extends semantic observations under geometric and semantic consistency constraints.

As illustrated in Fig. 1, Cross-View Reliability-Aware Aggregation estimates observation reliability from cross-view semantic consistency and combines it with rendering significance to construct a 3D semantic consensus. Consensus-Guided Feature Refinement uses this consensus to correct noisy 2D language features, forming a 3D→2D→3D refinement loop. Geometry-Aligned Semantic Propagation further exploits Gaussian geometry to extend semantic observations under geometric and semantic consistency constraints. The entire process requires no additional network training.

4.2 CROSS-VIEW RELIABILITY-AWARE AGGREGATION

Rendering significance quantifies an observation's role in image formation, but does not directly measure its semantic reliability. Since r_{iv} is not directly observable, a cross-view consistency score c_{iv} is used as its unsupervised proxy. The basic assumption is that reliable observations of the same 3D entity should remain semantically consistent across views.

Reliability-Aware Consensus Aggregation. For the language feature $f_{iv} \in \mathbb{R}^D$ of Gaussian i observed in view v, let $\hat{f}_{iv} = f_{iv}/\|f_{iv}\|_2$ denote its normalized feature. Let N_i denote the number of valid observations of Gaussian i. Observations with zero-norm features are excluded. For $N_i \geq 2$, the self-excluded mean cosine similarity is defined as

$$c_{iv}^{\text{raw}} = \frac{1}{N_i - 1} \sum_{u \neq v} \hat{f}_{iv}^{\mathsf{T}} \hat{f}_{iu}. \tag{1}$$

The current observation is excluded from its own reliability estimation. Before aggregation, the raw score is clipped as $c_{iv} = \max(c_{iv}^{\text{raw}}, 0)$. For $N_i = 1$, we set $c_{iv} = 1$, so the single observation retains its rendering-significance weight. The rendering significance and semantic reliability are jointly used to define the aggregation weight $w_{iv} = s_{iv}c_{iv}$. Here, s_{iv} is the rendering-significance score in the current view, and c_{iv} measures cross-view semantic reliability. Their combination suppresses inconsistent observations. The resulting Gaussian language representation is

$$F_i = \frac{\sum_v w_{iv}\hat{f}_{iv}}{\sum_v w_{iv} + \epsilon}. \tag{2}$$

The representation F_i therefore forms a 3D semantic consensus from reliable multi-view evidence.

Efficient Self-Excluded Consensus Computation. Define $S_i = \sum_{u=1}^{N_i} \hat{f}_{iu}$. Since $\hat{f}_{iv}^{\mathsf{T}} \hat{f}_{iv} = 1$, the same score can be computed as

$$c_{iv}^{\text{raw}} = \frac{\hat{f}_{iv}^{\mathsf{T}} S_i - 1}{N_i - 1}. \tag{3}$$

The result is clipped as above before aggregation.

4.3 Consensus-Guided Feature Refinement

Cross-View Reliability-Aware Aggregation suppresses inconsistent observations, but noise in the original 2D language features remains. The aggregated Gaussian features integrate reliable evidence from multiple views and therefore provide a stable cross-view semantic consensus. **Consensus-Guided Feature Refinement** uses this consensus to correct noisy 2D language features.

Consistency-Guided Teacher Construction. Based on the reliability-aware weights in Eq. (2), a semantic consistency score is defined for Gaussian i as

$$q_i = \frac{\sum_v w_{iv}}{\sum_v s_{iv} + \epsilon}. \tag{4}$$

A larger q_i indicates stronger cross-view semantic agreement. To prevent low-significance Gaussians from producing unreliable teacher features, the maximum rendering significance is defined as $M_i = \max_v s_{iv}$, and Gaussians satisfying $M_i < \tau_c$ are filtered out. Let $\mathcal{I}_c = \{i \mid M_i \geq \tau_c\}$ denote the retained Gaussian set. For pixel p in view v, the semantic consistency score q_i modulates the standard alpha-compositing contribution to construct the teacher feature map:

$$G_v(p) = \sum_{i \in \mathcal{I}_c} T_{iv}(p) \alpha_{iv}(p) q_i F_i. \tag{5}$$

Here, $T_{iv}(p)$ denotes the accumulated transmittance and $\alpha_{iv}(p)$ denotes the alpha contribution of Gaussian i at pixel p. The transmittance is recomputed over the retained Gaussians after pruning. The score q_i weights only the rendered feature and does not affect transmittance. The region-level teacher feature Z_{jv} is obtained by average-pooling $G_v(p)$ over the pixels in region j.

Consensus-Driven Adaptive Correction. Let X_{jv} denote the original 2D feature of region j in view v, and let Z_{jv} denote the corresponding teacher feature. If Z_{jv} has zero norm, correction is skipped and the original X_{jv} is retained. Otherwise, their ℓ_2-normalized representations are denoted by $\hat{X}_{jv}$ and $\hat{Z}_{jv}$. Their semantic consistency is defined as $\rho_{jv} = \hat{X}_{jv}^{\mathsf{T}} \hat{Z}_{jv}$. The correction strength is adaptively determined by

$$\eta_{jv} = \lambda_{\max} \operatorname{clip}\left(\frac{\tau - \rho_{jv}}{\tau - \rho_{\min}}, 0, 1 \right), \tag{6}$$

Here, $\rho_{\min}$ denotes the minimum similarity threshold. The clipping operation bounds the correction strength: features with high agreement receive little or no correction, while strongly inconsistent features are corrected up to the maximum strength $\lambda_{\max}$.

Given the adaptive correction strength η_{jv}, the refined 2D feature is

$$Y_{jv} = \operatorname{Norm}\left((1 - \eta_{jv}) \hat{X}_{jv} + \eta_{jv} \hat{Z}_{jv} \right). \tag{7}$$

Features that agree with the 3D consensus largely preserve their original semantics, while inconsistent observations receive stronger consensus guidance. The refined 2D features are then used to recompute the cross-view consistency scores and reliability-aware aggregation weights, followed by re-aggregation using Eq. (2) to obtain an updated 3D semantic consensus. This consensus can be projected back to the image space for the next refinement round, forming a 3D→2D→3D iterative loop.

4.4 GEOMETRY-ALIGNED SEMANTIC PROPAGATION

After improving the reliability of existing observations, **Geometry-Aligned Semantic Propagation** further exploits Gaussian geometry to extend reliable semantic evidence. For each anchor camera, a neighboring source view is first selected according to camera distance and viewing-direction consistency. An intermediate view is then constructed between the two real views. Existing real-view features are back-projected into 3D using depth and then reprojected to the intermediate view. Each real source view is reprojected separately, with a z-buffer retaining its nearest valid projection at each pixel. Only pixels with valid projections from both source views proceed to consistency filtering and feature fusion.

Consistency-Based Filtering. To suppress erroneous correspondences, the reprojected candidates are processed by depth-consistency filtering and semantic-consistency filtering. Let z_{reproj} be a candidate's depth in the intermediate-camera coordinates, and let z_{target} be the expected depth rendered from the Gaussians at the corresponding intermediate-view pixel. Depth consistency is defined as

$$e_z = \frac{|z_{\mathrm{reproj}} - z_{\mathrm{target}}|}{\max\left(|z_{\mathrm{target}}|, \epsilon_z\right)}. \tag{8}$$

Only candidates satisfying $e_z < \tau_z$ are retained. Semantic consistency is then evaluated between the corresponding language features from the two real source views, with a correspondence retained only when $\hat{f}_a^{\mathsf{T}} \hat{f}_b \geq \tau_f$.

Reliable Feature Fusion. Features passing both filters are fused according to the camera interpolation ratio r:

$$F_p = \mathrm{Norm}\left((1-r)\hat{f}_a + r\hat{f}_b\right). \tag{9}$$

The cross-view semantic consensus is determined only from real-view observations. The semantic consistency scores of propagated observations are computed only against this real-view consensus. Their rendering significance is multiplied by their semantic consistency scores to form aggregation weights. These weights are scaled by λ in the final aggregation. Propagated observations provide complementary semantic evidence without participating in consensus estimation.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and Metrics. The evaluation uses four LERF scenes, namely *figurines*, *ramen*, *teatime* and *waldo_kitchen*, together with ten ScanNet scenes (Dai et al., 2017). LERF mIoU evaluation follows LangSplat (Qin et al., 2024). ScanNet mIoU and mAcc evaluation under the 19, 15 and 10 class settings follows OpenGaussian (Wu et al., 2024).

Implementation Details. LERF evaluation uses a mask threshold of 0.5, feature levels 1–3 and a 3DGS checkpoint at iteration 30 000. The language features have $D = 512$ dimensions. For refinement, the Gaussian pruning threshold is $\tau_c = 0.1$, the similarity pivot is $\tau = 0.99$, the minimum similarity is $\rho_{\min} = 0.90$, and the maximum teacher-fusion weight is $\lambda_{\max} = 0.5$. Teacher rendering retains the weighted 512-dimensional numerator without denominator normalization. Refinement rounds use versioned Gaussian checkpoints for re-aggregation. For propagation, the number of intermediate views equals the number of real views. Orientation-aware camera pairing considers ten neighboring candidates and prefers those with a viewing-direction cosine similarity greater than 0.9. If none qualifies, the nearest neighbor is used. The interpolation ratio follows $\mathrm{Beta}(0.2, 0.2)$. The depth-consistency threshold is $\tau_z = 0.02$, and the semantic-consistency threshold is $\tau_f = 0.9$. The propagation contribution coefficient is $\lambda = 1$. The Ours results in Tables 1 and 2 are obtained by jointly aggregating R2-refined real-view features and features propagated from them through Geometry-Aligned Semantic Propagation.

5.2 Comparison with Existing Methods

Table 1 compares open-vocabulary semantic segmentation on the four LERF scenes, covering methods that require scene-specific training and training-free methods. Baseline results in Tables 1 and 2 are taken from their respective sources. Ours is evaluated in this work.

On LERF, ReliaSplat demonstrates competitive open-vocabulary segmentation performance, indicating that improving the reliability of input semantic observations can effectively enhance downstream 3D semantic understanding. Results on ScanNet further show that this advantage remains stable across different class settings, demonstrating the scene and class generalization of the proposed feature enhancement strategy.

Table 1: Open-vocabulary segmentation on LERF (mIoU). T-F indicates training-free status.

Method	Venue	T-F	Figurines	Ramen	Teatime	Waldo	Mean
Occam's	BMVC'25	✓	58.6	51.0	70.2	65.3	61.3
VALA	3DV'26	✓	59.9	51.5	70.2	65.1	61.7
THGS	ACMMM'25	✓	57.3	43.5	68.3	50.7	54.9
ReLaGS	CVPR'26	✓	**64.7**	51.2	**81.0**	60.6	**64.4**
LangSplat	CVPR'24	×	44.7	51.2	65.1	44.5	51.4
LangSplatV2	NeurIPS'25	×	56.4	51.8	72.2	59.1	59.9
3D VL-GS	ICLR'25	×	58.1	**61.4**	73.5	54.8	62.0
LaGa	ICML'25	×	64.1	55.6	70.9	65.6	64.0
GAGS	AAAI'26	×	53.6	46.8	60.3	55.8	54.1
Ours	–	✓	60.5	51.8	76.3	**68.4**	64.3

Table 2: Open-vocabulary semantic segmentation on ScanNet.

			19 cls		15 cls		10 cls	
Methods	Venue	T-F	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
THGS	ACMMM'25	✓	34.39	50.74	39.61	57.07	46.38	64.74
Occam's	BMVC'25	✓	31.93	48.93	34.25	53.71	45.16	64.39
VALA	3DV'26	✓	32.11	50.05	35.10	54.77	46.21	65.61
ReLaGS	CVPR'26	✓	32.35	44.52	40.04	**60.59**	47.17	66.08
LangSplat	CVPR'24	×	3.78	9.11	5.35	13.20	8.40	22.06
OpenGaussian	NeurIPS'24	×	24.73	41.54	30.13	48.25	38.29	55.19
LaGa	ICML'25	×	32.50	49.10	35.50	53.50	42.60	63.20
Ours	–	✓	**36.50**	**53.89**	**40.68**	59.97	**49.21**	**68.19**

5.3 Ablation Studies

Table 3 shows that R1 improves mIoU over R0 on both LERF and ScanNet. With further refinement, ScanNet mIoU increases slowly, while mean LERF performance remains nearly unchanged between R2 and R3. Most gains arise within the first one or two rounds, with limited benefit from additional iterations.

Table 3: Effect of Refinement Rounds (mIoU). All results use reliability-aware aggregation. R0 uses the original language features, while R1–R3 use features refined for one to three rounds, respectively.

	ScanNet			**LERF**				
Round	19 cls	15 cls	10 cls	figurines	ramen	teatime	Waldo	Mean
R0	35.52	38.60	47.00	59.90	50.61	74.19	66.66	62.84
R1	36.26	40.12	48.86	**60.60**	**51.57**	75.78	67.64	63.90
R2	36.35	40.50	49.05	60.37	**51.57**	76.20	68.33	**64.12**
R3	**36.53**	**40.79**	**49.36**	60.21	51.48	**76.31**	**68.46**	64.11

To further assess whether enhanced language features generalize across methods, Fig. 2 uses real-view language features produced after the first round (R1) to train LangSplat and LangSplatV2, respectively. Compared with models trained with the original language features, both methods produce

predictions that are more concentrated on the queried objects and have fewer erroneous responses in irrelevant regions. This indicates that the feature enhancement is not limited to ReliaSplat itself.



Figure 2: Qualitative localization results on LERF. From left to right, the pairs show LangSplat on *teatime* ("coffee") and *ramen* ("kamaboko"), then LangSplatV2 on *waldo_kitchen* ("knife"). Each pair compares predictions trained with original (left) and first-round (R1) refined real-view features (right), with other settings unchanged. Dashed boxes mark ground-truth regions, and red circles mark query points.

Table 4 evaluates the three components. In (A), s and c denote s_{iv} and c_{iv}, respectively. (B) fixes the self-excluded aggregation from (A) and uses one refinement round. In (C), R1 and R2 follow Table 3. The labels "neither," "depth," "cosine," and "both" indicate the active consistency filters.

Table 4: Ablation of the three components on LERF (mIoU).

	Setting	figurines	ramen	teatime	Waldo	Mean
(A) Cross-View Reliability-Aware Aggregation	significance only ($w = s$)	58.62	50.27	72.17	65.68	61.69
	consensus only ($w = c$)	57.02	45.76	**77.16**	65.27	61.30
	self-inclusive consensus ($w = s\,c'$)	**59.90**	**50.62**	74.18	66.63	62.83
	self-excluded consensus ($w = s\,c$)	**59.90**	50.61	74.19	**66.66**	**62.84**
(B) Consensus-Guided Feature Refinement	w/o M_i pruning, w/o q_i weighting	60.57	51.19	75.52	67.06	63.59
	w/ M_i pruning, w/o q_i weighting	60.24	50.77	75.44	67.09	63.39
	w/o M_i pruning, w/ q_i weighting	60.42	51.19	75.61	66.88	63.53
	w/ M_i pruning, w/ q_i weighting	**60.60**	**51.57**	**75.78**	**67.64**	**63.90**
(C) Geometry-Aligned Semantic Propagation	R1 Refined	60.60	51.57	75.78	67.64	63.90
	R1 Refined + Propagation (both)	**60.93**	**51.92**	75.90	67.99	64.19
	R1 Refined + Pseudo RGB Re-extraction	57.53	49.97	75.44	67.61	62.64
	R2 Refined	60.37	51.57	76.20	68.33	64.12
	R2 Refined + Propagation (neither)	60.28	51.71	76.28	68.11	64.10
	R2 Refined + Propagation (depth)	60.30	51.72	76.28	68.43	64.18
	R2 Refined + Propagation (cosine)	60.46	51.73	**76.30**	68.42	64.23
	R2 Refined + Propagation (both)	60.49	51.76	**76.30**	**68.44**	**64.25**

Cross-View Reliability-Aware Aggregation. c' denotes the consistency score computed with the current observation included. Table 4 shows that jointly using rendering significance and semantic reliability consistently improves aggregation quality, indicating that they provide complementary observation information. Similar self-inclusive and self-excluded results point to joint weighting as the main source of the gain. Self-exclusion avoids self-similarity bias in reliability estimation.

Consensus-Guided Feature Refinement. Table 4 compares M_i-based pruning and q_i weighting in one refinement round. Enabling either alone does not improve over using neither, whereas enabling both gives the best result in this group. This suggests that the two operations work together in teacher construction.

Geometry-Aligned Semantic Propagation. Table 4 shows that propagation with both filters improves mean mIoU at R1 and R2. At R2, either filter improves on propagation without these thresholds, and their combination gives the highest mean. Re-extracting features from Gaussian-rendered pseudo-view RGB images is less effective than propagating existing real-view features.

5.4 Cross-Framework Feature Generalization

With the official implementations and evaluation protocols of LangSplat, LangSplatV2, VALA, and THGS unchanged, Table 5 shows that replacing only their original input language features with

ReliaSplat's first-round (R1) refined real-view features improves mean mIoU in all four frameworks while leaving mean mAcc largely stable, indicating better input feature quality rather than adaptation to a specific aggregation rule.

Table 5: Cross-framework feature generalization experiment on LERF. Each scene and the mean report separate mIoU and mAcc values (%).

			Figurines		Ramen		Teatime		Waldo		Mean	
Method	T-F	Feature	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
LangSplat	×	Original	43.27	**75.00**	47.68	67.61	58.14	**86.44**	47.29	81.82	49.10	77.72
		refined	**48.13**	73.21	**51.17**	**73.24**	**64.86**	84.75	**51.28**	**90.91**	**53.86**	**80.53**
LangSplatV2	×	Original	58.89	78.57	50.38	73.24	70.98	93.22	63.81	86.36	61.02	82.85
		refined	**58.92**	78.57	**51.60**	73.24	**74.23**	93.22	**67.90**	86.36	**63.16**	82.85
VALA	✓	Original	**60.48**	**99.28**	35.41	**96.73**	71.59	**98.16**	53.91	95.17	55.35	**97.34**
		refined	58.60	98.93	**38.25**	96.64	**73.03**	98.15	**59.15**	**95.54**	**57.26**	97.32
THGS	✓	Original	**51.76**	**99.28**	32.36	95.10	81.30	98.87	55.88	96.81	55.32	97.52
		refined	49.98	99.27	**34.26**	**95.43**	81.30	**98.88**	**58.23**	**96.91**	**55.94**	**97.62**

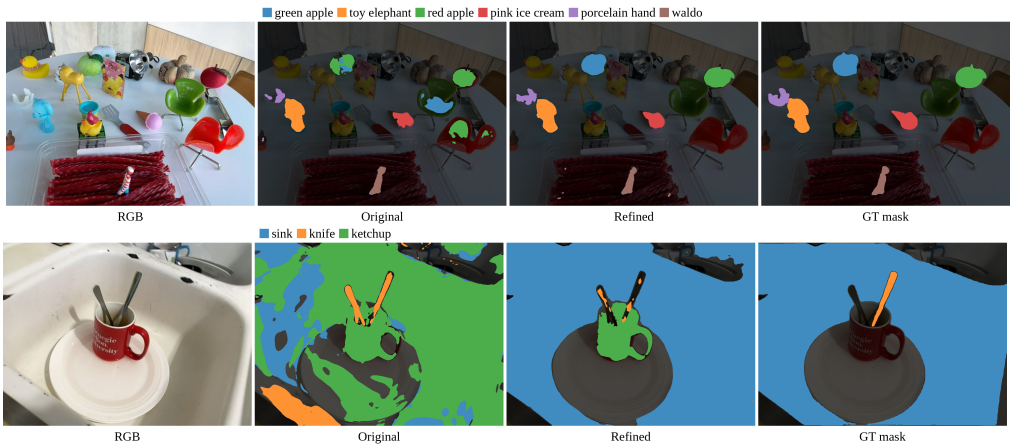


Figure 3: Qualitative mask comparison on LERF. The top row uses LangSplat on *figurines*, and the bottom row uses LangSplatV2 on *waldo_kitchen*. The legend lists the queried classes.

Figure 3 further provides qualitative evidence of cross-framework generalization. Refined features reduce erroneous responses between classes and make the predicted regions closer to the ground truth. This pattern is consistent with the quantitative results in Table 5.

6 Conclusion

ReliaSplat addresses unreliable input language features in Language Gaussian Splatting. Cross-View Reliability-Aware Aggregation identifies reliable semantic evidence through cross-view consistency. Consensus-Guided Feature Refinement uses this consensus to correct inconsistent 2D features. Geometry-Aligned Semantic Propagation complements the refined features with reliable real-view evidence. Experiments on LERF and ScanNet validate the framework in open-vocabulary semantic segmentation. Cross-framework results show that the enhanced features improve multiple Language Gaussian Splatting methods. These results indicate that ReliaSplat improves general input language-feature quality.

AI USE STATEMENT

Generative AI tools were used solely to assist with writing and language polishing. All AI-assisted text was reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Jiazhong Cen, Xudong Zhou, Jiemin Fang, Changsong Wen, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Tackling view-dependent semantics in 3D language gaussian splatting. In *Proceedings of the International Conference on Machine Learning*, 2025.

Jiahuan Cheng, Jan-Nico Zaech, Luc Van Gool, and Danda Pani Paudel. Occam's LGS: An efficient approach for language gaussian splatting. In *British Machine Vision Conference*, 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.

Shaohui Dai, Yansong Qu, Zheyan Li, Xinyang Li, Shengchuan Zhang, and Liujuan Cao. Training-free hierarchical scene understanding for gaussian splatting with superpoint graphs. In *Proceedings of the ACM International Conference on Multimedia*, pp. 3673–3682, 2025.

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

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, et al. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3D vision-language gaussian splatting. In *International Conference on Learning Representations*, 2025.

Yuning Peng, Haiping Wang, Yuan Liu, Chenglu Wen, Zhen Dong, and Bisheng Yang. GAGS: Granularity-aware feature distillation for language gaussian splatting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8376–8384, 2026. doi: 10.1609/aaai.v40i10.37787.

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, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Sen Wang, Kunyi Li, Siyun Liang, Elena Alegret, Jing Ma, Nassir Navab, and Stefano Gasperini. Visibility-aware language aggregation for open-vocabulary segmentation in 3D gaussian splatting. In *International Conference on 3D Vision*, 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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19114–19138, 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*, 2026.

A Supplementary Material

A.1 Derivation of Fast Self-Excluded Consensus

For a Gaussian with $N_i \geq 2$ valid observations and unit-normalised features, the direct pairwise definition in Eq. (1) can be rewritten by factoring the query observation out of the sum:

$$\begin{aligned} c_{iv}^{\text{raw}} &= \frac{1}{N_i - 1} \sum_{u \neq v} \hat{f}_{iv}^{\mathsf{T}} \hat{f}_{iu} \\ &= \frac{1}{N_i - 1} \hat{f}_{iv}^{\mathsf{T}} \left(\sum_{u=1}^{N_i} \hat{f}_{iu} - \hat{f}_{iv} \right) \\ &= \frac{\hat{f}_{iv}^{\mathsf{T}} S_i - \hat{f}_{iv}^{\mathsf{T}} \hat{f}_{iv}}{N_i - 1}. \end{aligned} \tag{A.1}$$

Because every observation is unit normalised, $\hat{f}_{iv}^{\mathsf{T}} \hat{f}_{iv} = 1$, and therefore

$$c_{iv}^{\text{raw}} = \frac{\hat{f}_{iv}^{\mathsf{T}} S_i - 1}{N_i - 1}. \tag{A.2}$$

This equality holds before nonnegative clipping. Applying $c_{iv} = \max(c_{iv}^{\text{raw}}, 0)$ to either form gives the same aggregation weight. Direct pairwise evaluation of all N_i scores costs $O(N_i^2 D)$. Computing S_i costs $O(N_i D)$ and all N_i dot products with S_i cost a further $O(N_i D)$. Thus, computing the cross-view consistency scores for Gaussian i takes $O(N_i D)$ time and does not allocate the $O(N_i^2)$ pairwise similarity matrix. This bound does not include visibility determination, rendering, or data movement in the full pipeline.

A.2 Per-Scene ScanNet Results

Tables A.1 and A.2 compare R2 refined real-view features with the same features augmented by Geometry-Aligned Semantic Propagation on ten ScanNet scenes. The two branches use the same Gaussian geometry and point-level evaluation protocol. Each scene uses level-3, 512-dimensional language features. Propagated observations are generated from R2 real-view features and re-aggregated with them. The means are equally weighted across scenes and computed from unrounded scores. All reported values are percentages.

Table A.1: Per-scene ScanNet mIoU (%). Each cell shows R2 refined real views → R2 with Geometry-Aligned Semantic Propagation.

Scene	19 cls	15 cls	10 cls
scene0000_00	39.07 → 38.80	37.57 → 37.50	44.30 → 44.44
scene0062_00	34.22 → 34.33	34.55 → 34.73	49.74 → 50.08
scene0070_00	32.57 → 32.61	41.68 → 41.98	53.38 → 53.76
scene0097_00	43.68 → 43.80	40.66 → 40.58	50.38 → 50.11
scene0140_00	23.73 → 24.07	34.18 → 34.61	43.71 → 44.36
scene0200_00	39.19 → 39.32	45.51 → 45.54	39.92 → 40.08
scene0347_00	45.07 → 45.05	44.61 → 44.39	50.22 → 50.03
scene0400_00	36.35 → 36.61	40.69 → 41.34	46.93 → 47.19
scene0590_00	38.09 → 38.38	47.17 → 47.21	67.25 → 67.29
scene0645_00	31.58 → 32.07	38.38 → 38.89	44.64 → 44.77
Mean	36.35 → 36.50	40.50 → 40.68	49.05 → 49.21

Table A.2: Per-scene ScanNet mAcc (%). The notation and evaluation protocol match Table A.1.

Scene	19 cls	15 cls	10 cls
scene0000_00	53.86 → 53.77	57.80 → 57.82	71.35 → 71.54
scene0062_00	66.39 → 66.41	67.06 → 67.15	79.00 → 79.21
scene0070_00	45.03 → 45.07	47.37 → 47.67	71.00 → 71.41
scene0097_00	63.74 → 63.75	69.17 → 69.02	74.36 → 74.14
scene0140_00	38.06 → 38.02	42.50 → 42.95	57.54 → 58.39
scene0200_00	49.84 → 49.93	60.52 → 60.33	54.36 → 54.46
scene0347_00	60.46 → 60.43	63.64 → 63.41	60.25 → 59.79
scene0400_00	57.30 → 57.99	65.85 → 66.74	68.32 → 68.89
scene0590_00	51.66 → 52.05	65.18 → 65.39	81.53 → 81.75
scene0645_00	51.01 → 51.50	58.74 → 59.23	62.32 → 62.36
Mean	53.73 → 53.89	59.78 → 59.97	68.00 → 68.19

Mean mIoU increases by 0.15, 0.18, and 0.17 percentage points under the 19-, 15-, and 10-class settings, respectively. Mean mAcc increases by 0.16, 0.19, and 0.19 points. The mIoU gains occur in 8, 7, and 8 of the ten scenes; the corresponding mAcc gains occur in 7, 7, and 8 scenes. Propagation uses a fixed camera schedule, while the scene-level changes vary.

A.3 Pseudo-Camera Pose Generation

For each scene, one pseudo-camera is constructed per real training camera. The training cameras are randomly permuted, and each is used once as an anchor A. Among the ten nearest neighbors in camera-center distance, candidates B are retained when their viewing-direction cosine exceeds 0.9. If none passes, the nearest neighbor is used as a fallback. The retained candidate with the highest score is selected, where

$$S(A,B) = \exp\left(-\frac{\|\mathbf{c}_A - \mathbf{c}_B\|_2}{\sigma}\right)\mathbf{v}_A^{\mathsf{T}}\mathbf{v}_B, \qquad \sigma = \mathrm{median}_A \min_{B\neq A}\|\mathbf{c}_A - \mathbf{c}_B\|_2. \tag{A.3}$$

Here, $\mathbf{c}$ is the camera center and $\mathbf{v}$ is the unit viewing direction. The scale σ is the median nearest-neighbor camera distance.

The Beta sampling and camera-pose interpolation are adapted from the camera-view blending strategy of 3D Vision-Language Gaussian Splatting (Peng et al., 2025). An interpolation ratio $r \sim \mathrm{Beta}(0.2, 0.2)$ is sampled. The camera center is interpolated linearly as $\mathbf{c}_p = (1-r)\mathbf{c}_A + r\mathbf{c}_B$, and camera rotations use quaternion spherical linear interpolation (Slerp). In this implementation, exact endpoint samples are rejected and fields of view are also interpolated linearly. The resulting camera-to-world transform is inverted to obtain the rendering pose. Each pseudo-camera is saved with its source pair and r. These fixed poses are reused across subsequent feature-generation and ablation runs.

Table A.3: LERF mIoU (%) under different interpolation distributions.

Interpolation distribution	Figurines	Ramen	Teatime	Waldo	Mean
Real-only (no pseudo-views)	60.60	51.57	75.78	67.64	63.90
$\mathrm{Beta}(0.1, 0.1)$	60.79	51.75	75.83	67.51	63.97
$\mathrm{Beta}(0.2, 0.2)$	**60.93**	**51.92**	**75.90**	**67.99**	**64.19**
$\mathrm{Beta}(0.3, 0.3)$	60.82	51.88	**75.90**	67.53	64.03
$\mathrm{Beta}(0.4, 0.4)$	60.82	51.73	75.85	67.64	64.01

Interpolation-Distribution Ablation on LERF. Table A.3 shows that the effect of the interpolation distribution varies across scenes. $\mathrm{Beta}(0.2, 0.2)$ has the strongest four-scene mean. It gives the highest scores on *figurines*, *ramen*, and *waldo_kitchen*, and ties for the highest score on *teatime*. Performance does not change monotonically with the distribution parameter. Some settings fall below the real-only result on *waldo_kitchen*.