QUERY-TIME CALIBRATION OF CLIP FEATURES FOR OPEN-VOCABULARY 3D SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Lifting vision-language features such as CLIP into 3D Gaussian reconstructions enables open-vocabulary segmentation through natural-language queries. Although CLIP represents a broad range of semantic concepts, the lifted features of individual scenes can exhibit strong directional concentration, while raw cosine similarity does not account for the scene-specific feature distribution. We show that substantial segmentation gains can be obtained by changing how a fixed lifted feature field is compared with text. Our method partially subtracts a shared scene direction from visual features and text prototypes and whitens text prototypes using a fixed reference vocabulary of generic terms. It then uses soft clustering for multiclass prediction and CSLS with a relevance margin against the reference vocabulary for single-query prediction. Beyond reducing directional concentration, scene centering improves class-relative feature separation in every evaluated scene. Across seven lifting methods and four datasets, the resulting query-time calibration consistently improves mean intersection over union in both multiclass and single-query segmentation. These gains require no retraining, source-image re-encoding, or feature re-lifting, demonstrating the value of scene-conditioned comparison for making better use of existing lifted representations.

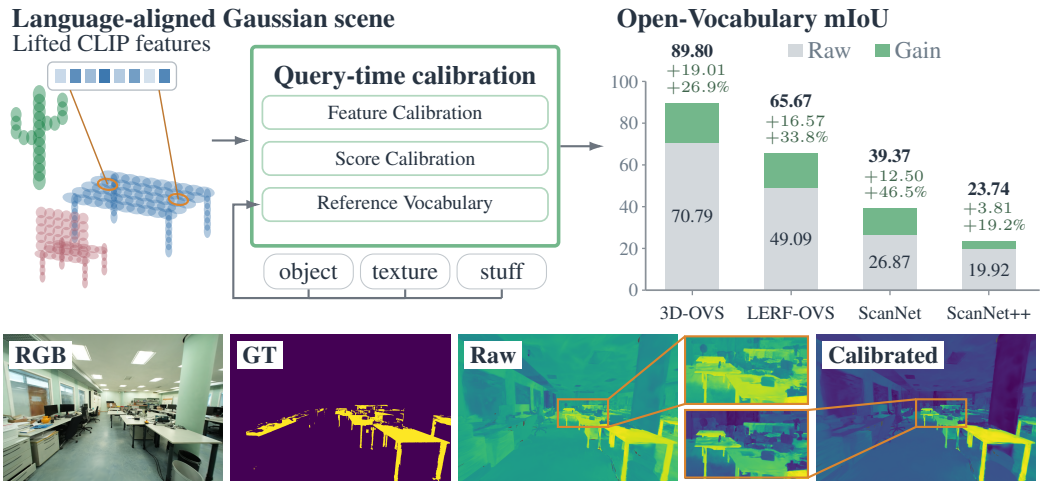


Figure 1: **Method Overview.** Starting from a Gaussian scene with lifted CLIP features, we calibrate the scene and text query features using centering and text-whitening, based on a reference vocabulary. Instead of plain query-scene similarities, we further calibrate the similarity scores using soft clustering for multi-class and CSLS with a relevance margin for single-query. Our calibration results in increased class-relative separation, as visualized in the similarity maps on the bottom. Our final predictions achieve substantial improvements in segmentation mIoU for both multi-class (visualized in bar chart) and single class.

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.

1 INTRODUCTION

High-quality 3D reconstructions provide persistent, navigable models of captured environments. Lifting vision–language features into them lets users locate and segment objects with natural-language queries rather than a closed classifier. CLIP supplies the shared image–text space (Radford et al., 2021), and methods like LERF (Kerr et al., 2023), LangSplat (Qin et al., 2024), and Feature 3DGS (Zhou et al., 2024) embed or distill language-aligned features into radiance fields and Gaussian representations. As reconstruction and feature lifting improve, comparing the resulting scene with text becomes an increasingly important part of open-vocabulary segmentation.

A particular reconstructed scene typically only presents a narrower semantic distribution than the broad image–text collection used to train CLIP. We find that its lifted features can share a pronounced scene-dependent direction, while raw cosine comparisons do not explicitly account for the actual distribution observed in that scene. This presents an opportunity to improve open-vocabulary queries solely by adapting them to a given scene at hand, rather than assuming that global embedding geometry is equally informative in every environment. We find that a substantial part of segmentation performance can be recovered at query time, through simple calibration rather than changes to reconstruction or feature lifting.

This scene-conditioned comparison should remain useful whether a user supplies a full class vocabulary or asks for a single object. Multiclass evaluation lets candidate classes compete for each region through an argmax: a region belongs to the class that scores highest among the supplied alternatives. In practice, however, a user may simply ask for a "fire extinguisher," without naming everything else in the scene. The task is then not to choose a winner among known alternatives, but to decide where the target is present and reject the remaining regions. Strong multiclass performance does not by itself establish this ability, motivating separate evaluation of single-query segmentation.

We address both settings through scene-conditioned calibration built around a fixed reference bank of generic terms. Our approach partially subtracts a shared scene direction from visual features and text prototypes and fits a regularized text-whitening transform solely to a scene-centered reference bank. For multiple requested classes, the calibrated similarities guide soft clustering of the primitive features similar to TransCLIP's assignment refinement (Zanella et al., 2024) without graph regularization. For a single query, we apply reduced CSLS (Lample et al., 2018) to the query and reference-bank scores. The reference bank then supplies negative competitors and we threshold the query's margin over their strongest reference competitor. This allows regions to be rejected without requiring the names of the scene's other classes. Our method improves comparison with already-lifted features without changing the reconstructed geometry, re-encoding source images, or re-lifting features.

Centering, whitening, and neighborhood-normalized similarity have established precedents (Mu & Viswanath, 2018; Su et al., 2021; Lample et al., 2018) in other domains, as does the use of generic negatives in language relevance scoring (Kerr et al., 2023). We show that these simple operations, adapted to the scene at hand, substantially improve open-vocabulary segmentation on existing lifted feature fields. The improvements hold across different lifting methods and extend to single-query segmentation, without retraining or re-lifting features. In summary, our contributions are:

- **Query-time improvements on fixed lifted fields.** We show that simple scene-conditioned calibration consistently improves open-vocabulary segmentation across seven lifting methods on four datasets, in both multiclass and single-query settings, without retraining, source-image re-encoding, or feature re-lifting.
- **Reference-bank calibration for multiclass and single-query segmentation.** We combine scene centering, reference-only text whitening, and scene-relative score adjustment in a query-time method. A fixed bank of generic terms defines the text transform independently of the requested class list and supplies negative competitors for single-query segmentation.
- **Analysis of feature geometry and calibration components.** We characterize directional concentration in lifted features and show that scene centering improves class-relative feature separation in every evaluated diagnostic scene. Component ablations examine how feature calibration, score adjustment, and reference vocabulary contribute to the final segmentation gains.

2 Related Work

Language-aware representations. Language-aware reconstruction methods determine which semantic information is available for subsequent queries. LERF (Kerr et al., 2023) embeds multiscale language features in a radiance field, while LangSplat (Qin et al., 2024) and Feature 3DGS (Zhou et al., 2024) distill features into Gaussian representations. LEGaussians (Shi et al., 2024) compresses language features through quantization, FMGS (Zuo et al., 2025) combines foundation-model features with a multiresolution feature field, and econSG (Zhang & Lee, 2025) addresses feature consistency across views. Gaussian Grouping (Ye et al., 2024), OpenGaussian (Wu et al., 2024), and N2F2 (Bhalgat et al., 2024) additionally organize instance identities or semantic scales. These approaches improve the 3D representation constructed from images.

Efficient lifting methods reduce the cost of constructing language-aware fields. Dr. Splat (Kim et al., 2025) directly registers language features, Occam's LGS (Cheng et al., 2025) aggregates multiview observations, and Splat Feature Solver (Xiong et al., 2026) formulates lifting as a regularized inverse problem. LightSplat (Bang et al., 2026) uses compact semantic indices and 3D clustering to associate image regions efficiently. Their feature-construction strategies are complementary to post-lifting calibration, which can operate on their outputs without repeating reconstruction or lifting.

Dense language features provide the inputs to many such pipelines. OpenSeg (Ghiasi et al., 2022) and LSeg (Li et al., 2022) learn dense visual-language alignment, while MaskCLIP (Zhou et al., 2022a) extracts dense predictions from CLIP (Radford et al., 2021). SAN (Xu et al., 2023) and OVSeg (Liang et al., 2023) adapt the segmentation representation through additional networks or masked-region training. SCLIP (Wang et al., 2025), ClearCLIP (Lan et al., 2025a), ProxyCLIP (Lan et al., 2025b), and CLIPtrase (Shao et al., 2025) instead rectify attention or token features without training. These methods are concerned with improving text and image features directly, rather than operating at query-time on already lifted fields.

Calibration also appears in dense and 3D segmentation. SCAN (Liu et al., 2024) learns proposal calibration and contextual adaptation, while ReCLIP (Wang & Kang, 2024) learns class and spatial bias corrections. OpenScene (Peng et al., 2023) fuses dense multiview features, and Open3DIS (Nguyen et al., 2024) associates image-guided proposals with language evidence. PLA (Ding et al., 2023), RegionPLC (Yang et al., 2024), and AIDE (Wang & Czarnecki, 2025) include learned confidence mechanisms for balancing known and novel classes. Our setting instead requires a label-free correction applicable to an already available field without learning a segmentation or confidence model.

Feature adaptation. Adapting embedding geometry provides one route to improving fixed-feature comparisons. Alignment and uniformity (Wang & Isola, 2020) characterize contrastive representations, while modality-gap analysis (Liang et al., 2022) and Two Effects, One Trigger (Schrodi et al., 2025) identify systematic differences between visual and text embeddings. All-but-the-Top (Mu & Viswanath, 2018) and isotropy analysis (Rajaee & Pilehvar, 2021) motivate removing dominant directions, while sentence whitening (Su et al., 2021) and Whitened CLIP (Betser et al., 2025) transform covariance structure. These works directly motivate our calibration components, with scene-local statistics and the reference vocabulary determining the calibration of a lifted field.

Other adaptations require evidence or optimization beyond stored scene features. CLIP Surgery (Li et al., 2025) removes image-conditioned redundant features, and CalibCLIP (Kang et al., 2025) modifies dominant-token processing. CoOp (Zhou et al., 2022c) and CoCoOp (Zhou et al., 2022b) learn prompts, TPT (Shu et al., 2022) and C-TPT (Yoon et al., 2024) optimize prompts at test time, and Tip-Adapter (Zhang et al., 2022) uses labeled support examples. ORION (Chakraborty et al., 2026) fine-tunes the text encoder to separate class prototypes. Their feature or prototype corrections target related errors, but their image access, labeled support, or learned parameter updates differ from our analytic transformation of frozen embeddings.

Distribution-conditioned adaptations are more direct competitors on fixed features. UMFC (Liang et al., 2024) estimates domain-dependent biases from unlabeled visual collections and corrects both visual and text features. PTC (Ma et al., 2026) builds visual prototypes from selected predictions and uses them to adjust text embeddings. Frolic (Zhu et al., 2024) constructs a covariance-based classifier, combines it with CLIP predictions, and iteratively corrects class priors. Applied to lifted primitives, these mechanisms compete with our scene centering and reference-text whitening by estimating corrections from the visual distribution or predicted class structure.

Score calibration. Reference-population normalization changes how fixed embeddings compete without modifying their encoders. Inverted softmax (Smith et al., 2017), CSLS (Lample et al., 2018), and querybank normalization (Bogolin et al., 2022) correct retrieval scores using statistics of other samples. Hubness (Radovanović et al., 2010) and prediction-concentration analysis (Sato et al., 2026) connect such corrections to unequal neighbor or class occurrence. Our multiclass branch uses soft clustering to refine primitive assignments, while our single-query branch applies class-side CSLS offsets before computing the query-versus-reference margin. Alternative scene-score normalizations are therefore direct comparisons for the single-query score correction.

Target-distribution methods instead adapt predictions jointly across samples or classes. BBSE (Lipton et al., 2018) and logit adjustment (Menon et al., 2021) connect prediction correction to class-frequency assumptions. OTTER (Shin et al., 2024) solves an optimal-transport assignment given an estimated target label distribution, while OnZeta (Qian & Hu, 2024) updates label statistics and visual-space class proxies online. TransCLIP (Zanella et al., 2024) iteratively estimates class distributions and propagates assignments over a feature graph. We adopt its soft assignment refinement without graph regularization for multiclass prediction and combine it with scene centering and reference-bank text whitening.

Single-query segmentation additionally requires a decision between foreground and background. LERF (Kerr et al., 2023) and FMGS (Zuo et al., 2025) compare a query against four canonical negative phrases, N2F2 (Bhalgat et al., 2024) uses generic phrases in multiscale feature learning, and NegLabel (Jiang et al., 2024) mines negative labels for out-of-distribution detection. GOI (Qu et al., 2024) replaces fixed-threshold selection with a query-specific semantic hyperplane optimized using a referring-expression model, while FreeCP (Chen et al., 2025) removes redundant candidate classes to reduce visual-language ambiguity. These approaches address negative competition, query-specific selection, or candidate-set quality.

3 METHOD

As outlined in Figure 1, our approach starts with a fixed Gaussian scene with lifted CLIP features and performs two steps. First we calibrate features using scene centering (Sec. 3.1) and text-queries are whitened (Sec. 3.2) using our reference vocabulary (Sec. 3.3). In the second step we calibrate per-query scores for both multi-class (Sec. 3.4) and single-query predictions (Sec. 3.5).

3.1 SCENE CENTERING

First, we calibrate scene and text features as illustrated in Figure 2. Let $\boldsymbol{x}_i, \boldsymbol{t}_c \in \mathbb{R}^d$ be unit-normalized column-vector features for valid primitives $i \in \mathcal{V}$ and text prototypes c, respectively. Both lie in the same CLIP space (Radford et al., 2021) that covers much broader semantics than the scene at hand. The calibration is independent of the lifting method, provided its output features share the text encoder's CLIP space. For a column vector $\boldsymbol{z} \in \mathbb{R}^d$, $\nu(\boldsymbol{z}) = \boldsymbol{z}/\max(\|\boldsymbol{z}\|_2, 10^{-12})$ denotes normalization, for matrices ν normalizes each row. The centering subtracts the shared scene direction, represented by the normalized mean $\boldsymbol{\mu} \in \mathbb{R}^d$:

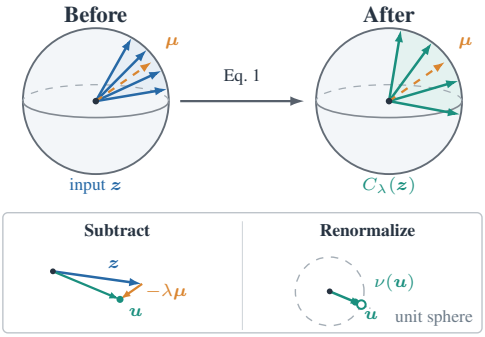


Figure 2: **Scene centering** maps blue input vectors $\boldsymbol{z}$ to teal calibrated vectors $\boldsymbol{u}$ by subtracting the orange mean direction $\boldsymbol{\mu}$ and renormalizing.

$$\boldsymbol{\mu} = \nu\left(\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \boldsymbol{x}_i\right), \qquad C_\lambda(\boldsymbol{z}) = \nu(\boldsymbol{z} - \lambda\boldsymbol{\mu}) \tag{1}$$

with $\lambda = 0.3$. For a row-stacked matrix $\boldsymbol{Z} \in \mathbb{R}^{r\times d}$, the row-wise extension is explicitly $C_\lambda(\boldsymbol{Z}) = \nu(\boldsymbol{Z} - \lambda \mathbf{1}_r \boldsymbol{\mu}^\top)$. Feature Calibration applies this centering step symmetrically, with $\widetilde{\boldsymbol{x}}_i = C_\lambda(\boldsymbol{x}_i)$ calibrating per-primitive features and $\widetilde{\boldsymbol{t}}_c = C_\lambda(\boldsymbol{t}_c)$ calibrating text query features.

3.2 TEXT-QUERY WHITENING

Next, we whiten query features using statistics fitted only to the scene-centered reference vocabulary. For a fixed scene, this whitening transform is independent of the requested class list. We scene-center the reference features and subtract their mean.

$$\boldsymbol{g}_b \in \mathbb{R}^d, \qquad \boldsymbol{G}_B = \begin{bmatrix} \boldsymbol{g}_1^\top \\ \vdots \\ \boldsymbol{g}_B^\top \end{bmatrix} \in \mathbb{R}^{B\times d}, \qquad \text{with } B = 45.$$

Here $\boldsymbol{g}_b$ is the unit text feature for entry b in reference-bank $\boldsymbol{G}_B$. Applying the centering gives

$$\widetilde{\boldsymbol{G}}_B = C_\lambda(\boldsymbol{G}_B), \qquad \bar{\boldsymbol{g}} = \frac{1}{B}\widetilde{\boldsymbol{G}}_B^\top \mathbf{1}_B, \qquad \boldsymbol{G}_0 = \widetilde{\boldsymbol{G}}_B - \mathbf{1}_B \bar{\boldsymbol{g}}^\top. \tag{2}$$

Here $\mathbf{1}_B \in \mathbb{R}^B$ is the all-ones vector. With the resulting mean $\bar{\boldsymbol{g}} \in \mathbb{R}^d$ of the centered reference bank $\widetilde{\boldsymbol{G}}_B$ and the residual matrix $\boldsymbol{G}_0 \in \mathbb{R}^{B\times d}$, the covariance is

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{G}_0^\top \boldsymbol{G}_0}{B-1} \in \mathbb{R}^{d\times d}. \tag{3}$$

Note that only the reference bank contributes to the fitted mean and covariance. We compute the eigenvector matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ to define

$$\boldsymbol{\Sigma}_\rho = \boldsymbol{\Sigma} + \rho \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d} \boldsymbol{I}_d, \qquad \boldsymbol{\Sigma}_\rho = \boldsymbol{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_d)\boldsymbol{U}^\top.$$

with $\rho = 10^{-3}$ and $\widehat{\sigma}_j = \max\big(\sigma_j, 10^{-10}\max_\ell \sigma_\ell\big)\,\forall j = 1, \ldots, d$ being used to stabilize subsequent inverse-power scaling when computing the whitening matrix $\boldsymbol{W}_{\alpha,B} \in \mathbb{R}^{d\times d}$:

$$\boldsymbol{W}_{\alpha,B} = \boldsymbol{U}\,\mathrm{diag}(\widehat{\sigma}_1^{-\alpha}, \ldots, \widehat{\sigma}_d^{-\alpha})\boldsymbol{U}^\top, \qquad \mathcal{W}_B(\boldsymbol{z}) = \nu(\boldsymbol{W}_{\alpha,B}(\boldsymbol{z} - \bar{\boldsymbol{g}}))\,. \tag{4}$$

with $\alpha = 0.40$ determined experimentally. Thus every query text feature is transformed as $\boldsymbol{t}'_c = \mathcal{W}_B(C_\lambda(\boldsymbol{t}_c))$. Similarly we define $\mathcal{W}_B(\boldsymbol{Z}) = \nu\big((\boldsymbol{Z} - \mathbf{1}_r\bar{\boldsymbol{g}}^\top)\boldsymbol{W}_{\alpha,B}^\top\big)$ for matrix $\boldsymbol{Z} \in \mathbb{R}^{r\times d}$.

Only the reference bank determines $\bar{\boldsymbol{g}}$ and $\boldsymbol{W}_{\alpha,B}$ and requested queries do not enter these statistics. Whitening is applied only to text features, while visual features remain scene-centered as $\widetilde{\boldsymbol{x}}_i = C_\lambda(\boldsymbol{x}_i)$ and are not whitened. $\mathcal{W}_B$ defines the whitening operation we subsequently apply to queries. We use regularized inverse-power scaling slightly weaker than inverse-square-root scaling typically employed in standard whitening. While queries are not used within whitening, our query calibration remains scene-conditioned through the use of scene-centered features.

3.3 REFERENCE VOCABULARY

The fixed vocabulary represents the text-side shared concepts using generic terms. Our bank G_B contains 45 reference unit text features $\boldsymbol{g}_b$ and includes only generic concepts like "object", "things", "texture", and further generic scene, material, and spatial descriptors, but no class names. Appendix B lists the complete bank and its text-encoding convention. As detailed in Sections 3.4 and 3.5, we use the reference vocabulary to determine whitening statistics for text query calibration and in the single-query case as background competitors.

To construct the vocabulary, we start from a fixed pool of 64 generic words and phrases and successively remove words deemed unhelpful. At each step, we test removing each entry and remove the word that gives the largest increase in mean mIoU on held-out calibration scenes, weighting the multi-class and single-query tasks and the scenes equally. We repeat this process until no removal improves this objective, leaving 45 entries, and then fix the bank.

3.4 MULTI-CLASS SCORE CALIBRATION

For $N > 1$ requested classes, we adopt the soft k-means-like refinement from TransCLIP (Zanella et al., 2024). Specifically, we apply the clustering to our calibrated features, but without their proposed similarity graph regularization. We alternate between soft class assignments and updates to

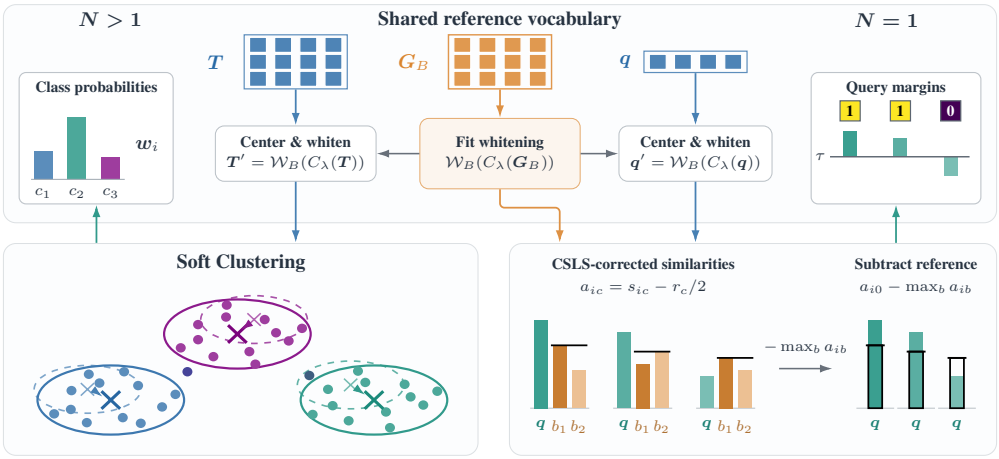


Figure 3: **Score calibration.** Starting from the **Middle:** First we use the reference vocabulary to determine the whitening statistics. **Left:** For multi-class queries ($N > 1$) we center and whiten all text queries and perform soft clustering of the primitive features, using fixed text-derived probabilities to guide assignments and weighted top-eight visual primitives to initialize the centers. After clustering, argmax decides the class of each primitive. **Right:** For a single-class query ($N = 1$) we center and whiten the query and the reference vocabulary, apply reduced CSLS and compute the margin between each primitive's corrected query score and its highest corrected reference-bank score. To make a prediction we threshold this relevance margin using τ.

the class centers and shared diagonal variance. We first compare each scene-centered visual feature $\widetilde{\boldsymbol{x}}_i$ with the calibrated text prototype $\boldsymbol{t}'_c$ for $c = 1, \ldots, N$:

$$s_{ic} = \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{t}'_c, \qquad \pi_{ic} = \frac{\exp(100 s_{ic})}{\sum_{c'=1}^{N} \exp(100 s_{ic'})}. \tag{5}$$

These initial class probabilities remain fixed and guide the refinement. For clustering, we initialize each class center from its eight highest-probability primitives, weighted by their probabilities and normalized to unit length. We then alternate between updating soft class assignments w_{ic} and estimating class centers $\boldsymbol{m}_c$ and a shared diagonal variance v_j from the assigned visual features. We update soft assignments using the initial probabilities and distances to the current class centers, then refit the centers and shared diagonal variance:

$$\begin{aligned} w_{ic} &\propto \pi_{ic} \exp\left(-\frac{1}{2\beta} \sum_{j=1}^{d} \frac{(\widetilde{x}_{ij} - m_{cj})^2}{v_j}\right), \qquad \sum_{c=1}^{N} w_{ic} = 1, \\ \boldsymbol{m}_c &\leftarrow \nu\left(\frac{\sum_{i \in \mathcal{V}} w_{ic} \widetilde{\boldsymbol{x}}_i}{\sum_{i \in \mathcal{V}} w_{ic}}\right), \qquad v_j \leftarrow \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \sum_{c=1}^{N} w_{ic} (\widetilde{x}_{ij} - m_{cj})^2. \end{aligned} \tag{6}$$

For each primitive, we normalize the assignment weights by dividing the right-hand side of the proportionality by its sum over all classes, so $\sum_{c=1}^{N} w_{ic} = 1$. We then update the class centers and variance in that order, with ν normalizing each center to unit length. Here v_j is the shared variance in feature dimension j, initialized to $1/d$. We use $\beta = 50$ and 11 iterations, following TransCLIP's implementation (Zanella et al., 2024). Finally, We predict $\widehat{y}_i = \arg\max_{1 \le c \le N} w_{ic}$.

3.5 Single-Query Score Calibration

With one requested query $\boldsymbol{q}$, a class argmax alone cannot reject background and we cannot perform soft clustering with a single prototype. Instead, we calibrate the query's similarity scores using CSLS (Lample et al., 2018) and compare them against the most similar prototype in our reference bank $\boldsymbol{G}_B$.

We apply the scene centering and fixed reference-bank whitening transform to both the query and the reference words. For each resulting prototype $\boldsymbol{t}'_c$, where $c \in \{0, \ldots, B\}$ with $\boldsymbol{t}'_0 = \boldsymbol{q}'$, we compute its similarity to the scene features $s_{ic} = \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{t}'_c$. Then we take the average of the top-1000

similarities per class r_c, which is used to correct each score: $a_{ic} = s_{ic} - \frac{1}{2} r_c$. This reduced CSLS correction adjusts each query's score baseline while preserving its ranking of primitives.

Let a_{i0} denote the corrected query score and a_{ib} the score for reference word b. We subtract the strongest reference score to obtain the *relevance margin*:

$$R_i(\boldsymbol{q}) = a_{i0} - \max_{1 \le b \le B} a_{ib}, \qquad \widehat{M}_i(\boldsymbol{q}) = \mathbf{1}[R_i(\boldsymbol{q}) \ge \tau]\,. \tag{7}$$

The indicator $\mathbf{1}[\cdot]$ is one when its condition holds and zero otherwise. The margin measures whether the requested query scores above the generic reference words at that primitive. We use one threshold τ across all scenes and queries within each dataset and lifted feature set.

4 Evaluation

We evaluate our calibration on its ability to increase class-relative separation, 3D open-vocabulary segmentation and present ablation studies. For the open-vocabulary segmentation we use LERF-OVS Qin et al. (2024), 3D-OVS Liu et al. (2023), ten scenes from ScanNet (Dai et al., 2017) and twelve from ScanNet++ (Yeshwanth et al., 2023) using mean intersection-over-union (mIoU), precision and recall. We evaluate lifted features from SFS (Xiong et al., 2026), NormLift (Zang et al., 2026), LUDVIG (Marrie et al., 2025), OpenGaussian (Wu et al., 2024), LangSplat (Qin et al., 2024), Occam (Cheng et al., 2025) and VALA (Wang et al., 2026). To assign ground truth points to Gaussians, we follow the respective methods' strategies for discarding low-opacity and invalidly lifted Gaussians, and assign the remaining eligible Gaussians using minimum Mahalanobis distance. Appendix E presents more details on our evaluation protocol.

In Section 4.2 we compare against raw uncalibrated features, as well as previous work on CLIP calibration. These methods were not designed for our Gaussian-lifted 3D task, but we adopt their ideas to enable comparison. Specifically we compare to adaptations of TransCLIP (Zanella et al., 2024), PTC (Ma et al., 2026), UMFC (Liang et al., 2024) and a Frolic (Zhu et al., 2024) inspired baseline.

4.1 Investigating the Effects of Feature Calibration

We quantify the directional concentration illustrated in Figure 2 using the von Mises–Fisher (vMF) concentration estimate κ. Higher values indicate stronger concentration around a common direction. As intended, subtracting the shared scene direction reduces mean κ from 1600 to 517 across the twelve ScanNet++ SFS scenes. The uncentered second moment averages feature outer products without subtracting their mean, capturing both a shared direction and the spread around it. Its leading eigenvalue's share measures the fraction of total squared feature magnitude along the strongest axis, this falls from 76.8% to 47.3%. We also analyze participation rank, which summarizes how many axes effectively carry this magnitude: k equally weighted axes give a participation rank of k. It rises from 1.69 to 4.15, indicating that the features are less dominated by a few directions. Together, these descriptive statistics show that centering redistributes feature magnitude across more axes, rather than merely reducing the mean direction.

To assess semantic discrimination, we group features by their GT class and normalize each class's average feature to obtain its mean direction. Between-class scatter is the average squared distance between these class mean directions: it measures how far apart the class centers are. Within-class scatter is the average squared distance of features from their own class's mean direction, averaged equally across classes: it measures how spread out each class is around its center. Their ratio measures separation between classes relative to their internal variation. The mean class-relative separation ratio increases from 1.820 to 2.056 on ScanNet (19 classes), from 2.854 to 3.163 on ScanNet++ (100 classes), and from 2.414 to 2.738 on ScanNet++ (20 classes). The ratio increases in every scene in all three settings. We conclude that the transformation does not merely spread feature directions more broadly, it improves their separation relative to the ground-truth semantic classes.

4.2 Open-Vocabulary Segmentation

Multi-Class Calibration. We present our improvements for multi-class evaluation in the top half of Table 1. We report LERF-OVS, 3D-OVS, ScanNet with the 19-class targets and ScanNet++

Table 1: **Segmentation Results Raw vs. Calibrated**. Top: multi-class ($N > 1$). Bottom: single-query ($N = 1$). The $N = 1$ results use the fixed held-out thresholds stated in Appendix C. Smaller ScanNet 10-class and ScanNet++ 20-class results, as well as precision and recall are reported in Appendix A.

	LERF-OVS			3D-OVS			ScanNet (19-class)			ScanNet++ (100-class)		
Method	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ
Multi-Query $N > 1$												
SFS Xiong et al. (2026)	49.09	**65.67**	+16.57	70.79	**89.80**	+19.01	26.87	**39.37**	+12.50	19.92	**23.74**	+3.81
NormLift Zang et al. (2026)	58.55	**60.25**	+1.70	70.90	**86.11**	+15.20	31.40	**39.55**	+8.14	20.60	**24.76**	+4.16
LUDVIG* Marrie et al. (2025)	38.05	**45.44**	+7.39	4.15	**5.85**	+1.70	29.32	**39.83**	+10.51	20.31	**23.96**	+3.66
Occam Cheng et al. (2025)	57.13	**61.05**	+3.92	66.19	**84.21**	+18.01	23.87	**32.11**	+8.24	20.18	**23.16**	+2.98
VALA Wang et al. (2026)	**59.83**	59.68	−0.15	64.16	**79.02**	+14.86	27.81	**36.16**	+8.35	19.28	**22.90**	+3.61
OpenGaussian Wu et al. (2024)	33.62	**46.18**	+12.56	23.73	**28.81**	+5.08	24.34	**32.09**	+7.76	15.25	**17.72**	+2.47
LangSplat Qin et al. (2024)	1.49	**5.15**	+3.66	7.43	**15.19**	+7.76	3.04	**4.00**	+0.96	4.50	**6.08**	+1.58
Single-Query $N = 1$												
SFS Xiong et al. (2026)	45.71	**53.86**	+8.15	69.21	**83.59**	+14.39	26.66	**34.52**	+7.85	22.55	**24.39**	+1.84
NormLift Zang et al. (2026)	48.82	**53.03**	+4.21	68.62	**82.47**	+13.86	30.56	**36.64**	+6.08	23.89	**25.54**	+1.65
LUDVIG* Marrie et al. (2025)	31.98	**35.15**	+3.16	13.94	**13.96**	+0.03	28.46	**35.76**	+7.30	23.11	**25.27**	+2.16
Occam Cheng et al. (2025)	46.09	**53.67**	+7.58	61.10	**75.28**	+14.17	22.96	**31.01**	+8.05	22.97	**24.24**	+1.27
VALA Wang et al. (2026)	21.30	**54.15**	+32.85	31.19	**75.01**	+43.82	26.60	**32.85**	+6.26	21.99	**22.67**	+0.68
OpenGaussian Wu et al. (2024)	30.97	**31.06**	+0.09	30.72	**34.24**	+3.52	22.95	**30.81**	+7.86	18.91	**22.08**	+3.17
LangSplat Qin et al. (2024)	4.76	**5.19**	+0.43	14.96	**19.57**	+4.61	11.37	**11.62**	+0.25	5.92	**8.05**	+2.13

* We could not reproduce reasonable 3D-OVS results for LUDVIG.

with the 100-class targets. The smaller ScanNet 10-class and ScanNet++ 20-class vocabularies can be found in Appendix A. The results show that our calibration consistently improves mIoU in all datasets and across all methods, except for VALA and OpenGaussian on LERF-OVS, which are mostly unchanged. We find that our calibration can sometimes shift the precision/recall trade-off compared to the uncalibrated baseline (see Appendix A), but in most cases we improve all metrics.

Single-Query Calibration. In the bottom half of Table 1 we present results for our calibration when using only a single query at a time. We query all labels present in a scene individually and average over present classes, then over scenes. Unlike the multi-class ablation, these $N = 1$ results threshold one query's target-minus-bank margin rather than taking an argmax over the real class vocabulary. The raw similarity and calibrated relevance-margin thresholds τ were determined on held-out scenes and then fixed for all evaluation scenes and queries. The dataset-specific values and LangSplat exceptions are stated in Appendix C.

The resulting segmentation results are generally below their multi-class counterparts. This pattern is consistent with the absence of mutual exclusion between independently thresholded query masks. As a result different queries could re-use the same primitives that are otherwise competed for in the multi-class case. Nevertheless, our calibration improves IoU in all cases, especially for VALA on LERF-OVS. To investigate why raw performs so poorly in that case, we select an oracle raw-score threshold on each evaluation scene, yielding only 23.71% mIoU, barely improving over the held-out threshold. It seems that sparse semantic support attenuates rendered scores differently across pixels, affecting absolute thresholding but not class-relative argmax decisions. Qualitative results are available in Appendix A.

4.3 Comparison to other Query-Time Adaptations

In Table 2 we compare our method to four other inference-only adaptations using the fixed SFS feature fields. Specifically, we adapt PTC (Ma et al., 2026) by replacing image tokens with valid scene primitives. High-margin initial predictions form class-specific visual prototypes that interpolate the text embeddings. We apply this once per scene, without crop-specific processing, and retain the released defaults ($\rho = 0.1$, $K_{\min} = 40$, $\mu = 0.1$). Our Frolic-inspired baseline (Zhu et al., 2024) estimates covariance from scene features and text prototypes, using shrinkage of 0.9. It removes the Gaussian intercept and normalizes the transformed

Table 2: **Comparison to other adaptations.** Values are mIoU (Δ over raw) for $N > 1$ on SFS. Other methods are unsuitable for $N = 1$.

Method	ScanNet	ScanNet++
Ours	**39.37** (+12.50)	**23.74** (+3.81)
TransCLIP	35.54 (+8.68)	22.81 (+2.89)
PTC	27.92 (+1.05)	20.81 (+0.89)
Frolic-inspired	28.37 (+1.51)	16.90 (−3.02)
UMFC	11.31 (−15.55)	9.23 (−10.69)

classifier, rather than reproducing full Frolic's fusion and iterative prior correction. We also adapt UMFC (Liang et al., 2024) by clustering valid scene features into six groups and applying the resulting cluster-based corrections to visual features and text prototypes. We retain UMFC's released correction strengths of 0.9 and 0.3. Lastly, we adapt TransCLIP (Zanella et al., 2024) to each scene by treating its lifted Gaussian features as unlabeled samples and performing iterative refinement similar to subsection 3.4, but instead of the restricted adaptation we use, this baseline employs full TransCLIP with a feature-similarity neighbor graph.

We evaluate these adaptations only for multiclass segmentation, as they require multi-class queries and are inherently unsuited for single-query calibration. The results show that PTC and TransCLIP variants can also improve segmentation performance, while the Frolic variant only improves on ScanNet. Our full calibration outperforms these adaptations, showing that additional feature calibration and reference vocabulary are complementary to the soft clustering idea of TransCLIP.

4.4 Contribution ablations

Table 3 evaluates the three conceptual contributions used by the full $N > 1$ method. Feature Calibration combines symmetric centering and regularized text whitening, Score Calibration uses soft clustering, and Reference Vocabulary enables our bank of generic words. When Reference Vocabulary is turned off, we use the $N > 1$ queries as the whitening basis. The full method yields the largest mIoU gain on both datasets. On ScanNet++ Feature Calibration alone has a negative impact, while being helpful in every combination with other components. In Appendix D we show additional experiments on sensitivity of λ and α parameters.

Table 3: **Component ablation.** Values are mIoU (Δ over raw) using SFS.

C1	C2	C3	ScanNet		ScanNet++	
			26.87		19.92	
✓			31.83	(+4.97)	18.77	(−1.15)
	✓		35.09	(+8.22)	23.66	(+3.73)
✓	✓		36.37	(+9.50)	21.80	(+1.88)
✓		✓	36.49	(+9.63)	21.15	(+1.23)
✓	✓	✓	**39.37**	(+12.50)	**23.74**	(+3.81)

C1 = Feature Calibration, C2 = Score Calibration, C3 = Reference Vocabulary

5 Conclusion

We studied scene-conditioned calibration for open-vocabulary 3D segmentation while holding reconstructed geometry and lifted features fixed. Combining scene centering, regularized text-prototype whitening, and scene-relative score adjustment improves mIoU across all evaluated lifting methods on four datasets, in both multiclass and single-query settings. Geometric diagnostics show that centering improves class-relative feature separation, while component ablations demonstrate the importance of combining feature, score calibration and our fixed reference vocabulary to determine text-whitening statistics. Multiclass prediction uses soft clustering, while single-query prediction uses a reduced-CSLS margin against the reference vocabulary. These results establish query-time calibration as a useful complement to feature lifting: segmentation can improve without retraining, re-encoding source images, or re-lifting features.

Limitations. Our evaluation is limited to indoor scenes, and the shared centering and whitening parameters were selected using SFS results on held-out indoor scenes, leaving transfer to other environments and embedding models untested. Single-query evaluation uses GT-present classes and held-out-selected thresholds, so it does not establish reliable rejection of entirely absent targets or arbitrary natural-language queries. Predictions depend on the requested vocabulary and reference bank, whose suitability across different scenes and queries is not established. Because calibration leaves geometry and feature lifting unchanged, it cannot reconstruct missing surfaces or recover semantic information absent from the lifted field.

Open Questions. For single-query segmentation, an especially promising question is how to choose negative queries that reflect the available scene context rather than relying only on a fixed generic bank. Previously queried concepts, when appropriate as negative alternatives, or prior knowledge of the scene could inform the reference vocabulary and provide more relevant competitors for the current target. Growing the reference vocabulary over time as a user queries the scene, especially when applied in similar environments, is an interesting future direction. Further, all tested scenes are generally in an indoor environment and evaluation in outdoor environments or substantially different scenes remains untested.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing sections of the paper. We also used AI tools to find related works beyond what our manual research found. We used AI models to review and critique our paper and used its feedback to improve our work. AI coding agents were used for parts of the implementation and debugging. All generated text, code and suggested papers were reviewed thoroughly by the authors. In all cases we used AI tools as *tools* to assist our work, never to replace or automate full aspects of the research process.

REPRODUCIBILITY STATEMENT

With our paper we release the full source code required to reproduce our results. We will upload the reconstructed 3DGS scenes and the lifted feature sets used as input to our calibration. We will supply a simple Python script that reproduces the exact segmentations that lead to the results reported in Table 1.

REFERENCES

Jaehun Bang, Jinhyeok Kim, Minji Kim, Seungheon Jeong, and Kyungdon Joo. LightSplat: Fast and memory-efficient open-vocabulary 3D scene understanding in five seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19812–19821, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bang_LightSplat_Fast_and_Memory-Efficient_Open-Vocabulary_3D_Scene_Understanding_in_Five_CVPR_2026_paper.html.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened CLIP as a Likelihood Surrogate of Images and Captions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 4069–4095. PMLR, 2025. URL https://proceedings.mlr.press/v267/betser25a.html.

Yash Bhalgat, Iro Laina, João F. Henriques, Andrew Zisserman, and Andrea Vedaldi. N2F2: Hierarchical Scene Understanding with Nested Neural Feature Fields. In *Computer Vision – ECCV 2024*, 2024. URL https://arxiv.org/abs/2403.10997.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5194–5205, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Bogolin_Cross_Modal_Retrieval_With_Querybank_Normalisation_CVPR_2022_paper.html.

Omprakash Chakraborty, Jose Dolz, and Ismail Ben Ayed. ORION: ORthonormal text encoding for universal VLM adaptatION. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31556–31565, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chakraborty_ORION_ORthonormal_Text_Encoding_for_Universal_VLM_AdaptatION_CVPR_2026_paper.html.

Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23124–23134, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chen_Training-Free_Class_Purification_for_Open-Vocabulary_Semantic_Segmentation_ICCV_2025_paper.html.

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. URL https://arxiv.org/abs/2412.01807.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-driven open-vocabulary 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7010–7019, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ding_PLA_Language-Driven_Open-Vocabulary_3D_Scene_Understanding_CVPR_2023_paper.html.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision – ECCV 2022*, pp. 540–557. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-20059-5_31. URL https://doi.org/10.1007/978-3-031-20059-5_31.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *International Conference on Learning Representations*, pp. 14913–14941, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/40eff1670d6b08bb1bda48b0c5f30110-Abstract-Conference.html.

Bin Kang, Bin Chen, Junjie Wang, Yulin Li, Junzhi Zhao, Junle Wang, and Zhuotao Tian. CalibCLIP: Contextual Calibration of Dominant Semantics for Text-Driven Image Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. doi: 10.1145/3746027.3755765. URL https://arxiv.org/abs/2510.05586.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023. URL https://arxiv.org/abs/2303.09553.

Jun-Seong Kim, GeonU Kim, Yu-Ji Kim, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. Splat: Directly Referring 3D Gaussian Splatting via Direct Language Embedding Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14137–14146, 2025. URL https://arxiv.org/abs/2502.16652.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *Computer Vision – ECCV 2024*, pp. 143–160. Springer, 2025a. doi: 10.1007/978-3-031-72970-6_9. URL https://doi.org/10.1007/978-3-031-72970-6_9.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *Computer Vision – ECCV 2024*, pp. 70–88. Springer, 2025b. doi: 10.1007/978-3-031-73113-6_5. URL https://doi.org/10.1007/978-3-031-73113-6_5.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RriDjddCLN.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A Closer Look at the Explainability of Contrastive Language-Image Pre-training. *Pattern Recognition*, 162:111409, 2025. doi: 10.1016/j.patcog.2025.111409. URL https://arxiv.org/abs/2304.05653.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7070, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Liang_Open-Vocabulary_Semantic_Segmentation_With_Mask-Adapted_CLIP_CVPR_2023_paper.html.

Jiachen Liang, Ruibing Hou, Minyang Hu, Hong Chang, Shiguang Shan, and Xilin Chen. UMFC: Unsupervised multi-domain feature calibration for vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 114072–114093. Curran Associates, Inc., 2024. doi: 10.52202/079017-3623. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cf026564af33fed49b911b9532680b9d-Paper-Conference.pdf.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625. Curran Associates, Inc., 2022. doi: 10.52202/068431-1280. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

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

Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3491–3500, June 2024.

Wanli Ma, Jiangwen Lu, Qinmu Peng, and Xinge You. Perceptual Anchoring: Prototype-Guided Text Calibration for Training-free Open-Vocabulary Semantic Segmentation. arXiv preprint, 2026. URL https://arxiv.org/abs/2608.03991.

Juliette Marrie, Romain Menegaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7440–7450, 2025. doi: 10.1109/ICCV51701.2025.00698.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.07314.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4018–4028, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Nguyen_Open3DIS_Open-Vocabulary_3D_Instance_Segmentation_with_2D_Mask_Guidance_CVPR_2024_paper.html.

Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Peng_OpenScene_3D_Scene_Understanding_With_Open_Vocabularies_CVPR_2023_paper.html.

Qi Qian and Juhua Hu. Online zero-shot classification with CLIP. In *Computer Vision – ECCV 2024*, pp. 462–477. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72980-5_27. URL https://doi.org/10.1007/978-3-031-72980-5_27.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D Language Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024. URL https://arxiv.org/abs/2312.16084.

Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. GOI: Find 3D gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5328–5337, 2024. doi: 10.1145/3664647.3680852. URL https://doi.org/10.1145/3664647.3680852.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531, 2010. URL https://jmlr.org/papers/v11/radovanovic10a.html.

Sara Rajaee and Mohammad Taher Pilehvar. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 575–584. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-short.73. URL https://aclanthology.org/2021.acl-short.73/.

Shota Sato, Hajime Kiyama, Tosho Hirasawa, and Mamoru Komachi. When Modality Gap Reduction Fails: Prediction-Level Hubness in CLIP. arXiv preprint, 2026. URL https://arxiv.org/abs/2609.01103.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uAFHCZRmXk.

Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of CLIP for training-free open vocabulary semantic segmentation. In *Computer Vision – ECCV 2024*, pp. 139–156. Springer, 2025. doi: 10.1007/978-3-031-73016-0_9. URL https://doi.org/10.1007/978-3-031-73016-0_9.

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 (CVPR)*, pp. 5333–5343, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Shi_Language_Embedded_3D_Gaussians_for_Open-Vocabulary_Scene_Understanding_CVPR_2024_paper.html.

Changho Shin, Jitian Zhao, Sonia Cromp, Harit Vishwakarma, and Frederic Sala. OTTER: Effortless label distribution adaptation of zero-shot models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44064–44101. Curran Associates, Inc., 2024. doi: 10.52202/079017-1399. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4df1cc5a7528b7197ad8ae76ff30107a-Paper-Conference.pdf.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14274–14289. Curran Associates, Inc., 2022. doi: 10.52202/068431-1038. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5bf2b802e24106064dc547ae9283bb0c-Abstract-Conference.html.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1Aab85gg.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval, 2021. URL https://arxiv.org/abs/2103.15316.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *Computer Vision – ECCV 2024*, pp. 315–332. Springer, 2025. doi: 10.1007/978-3-031-72664-4_18. URL https://doi.org/10.1007/978-3-031-72664-4_18.

Jingyun Wang and Guoliang Kang. Learn to rectify the bias of CLIP for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4102–4112, June 2024.

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 *2026 International Conference on 3D Vision (3DV)*, pp. 1791–1800, 2026. doi: 10.1109/3DV69130.2026.00170.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Yimu Wang and Krzysztof Czarnecki. AIDE: Improving 3D Open-Vocabulary Semantic Segmentation by Aligned Vision-Language Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/papers/Wang_AIDE_Improving_3D_Open-Vocabulary_Semantic_Segmentation_by_Aligned_Vision-Language_Learning_WACV_2025_paper.pdf.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. OpenGaussian: Towards Point-Level 3D Gaussian-based Open Vocabulary Understanding. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.02058.

Butian Xiong, Rong Liu, Kenneth Xu, Meida Chen, and Andrew Feng. Splat Feature Solver. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.12216.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2945–2954, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Xu_Side_Adapter_Network_for_Open-Vocabulary_Semantic_Segmentation_CVPR_2023_paper.html.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19823–19832, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_RegionPLC_Regional_Point-Language_Contrastive_Learning_for_Open-World_3D_Scene_Understanding_CVPR_2024_paper.html.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian Grouping: Segment and Edit Anything in 3D Scenes. In *Computer Vision – ECCV 2024*, 2024. URL https://arxiv.org/abs/2312.00732.

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

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, pp. 8636–8660, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/23ff02034404b65776080cbf7148addd-Abstract-Conference.html.

Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting vision-language models with transduction. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62223–62256. Curran Associates, Inc., 2024. doi: 10.52202/079017-1988. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/71d7dbe2652bd4662d29fa269f059db4-Paper-Conference.pdf.

Yihan Zang, Da Li, Dominik Engel, Shinkyu Park, and Ivan Viola. Normlift: From lifted features to semantic reliability in 3d gaussian splatting, 2026. URL https://arxiv.org/abs/2609.18898.

Can Zhang and Gim Hee Lee. econSG: Efficient and multi-view consistent open-vocabulary 3D semantic gaussians. In *International Conference on Learning Representations*, pp. 68368–68385, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ab3f114401f0523ca1cc09de0621f400-Paper-Conference.pdf.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*, pp. 493–510. Springer, 2022. doi: 10.1007/978-3-031-19833-5_29. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/154_ECCV_2022_paper.php.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *Computer Vision – ECCV 2022*, pp. 696–712. Springer Nature Switzerland, 2022a. doi: 10.1007/978-3-031-19815-1_40. URL https://doi.org/10.1007/978-3-031-19815-1_40.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, June 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhou_Conditional_Prompt_Learning_for_Vision-Language_Models_CVPR_2022_paper.html.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022c. doi: 10.1007/s11263-022-01653-1. URL https://link.springer.com/article/10.1007/s11263-022-01653-1.

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. URL https://arxiv.org/abs/2312.03203.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2001–2025. Curran Associates, Inc., 2024. doi: 10.52202/079017-0064. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/03cdf8e212ba92a3f36bffe1391928bd-Abstract-Conference.html.

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. FMGS: Foundation model embedded 3D gaussian splatting for holistic 3D scene understanding. *International Journal of Computer Vision*, 133(2):611–627, 2025. doi: 10.1007/s11263-024-02183-8. URL https://doi.org/10.1007/s11263-024-02183-8.

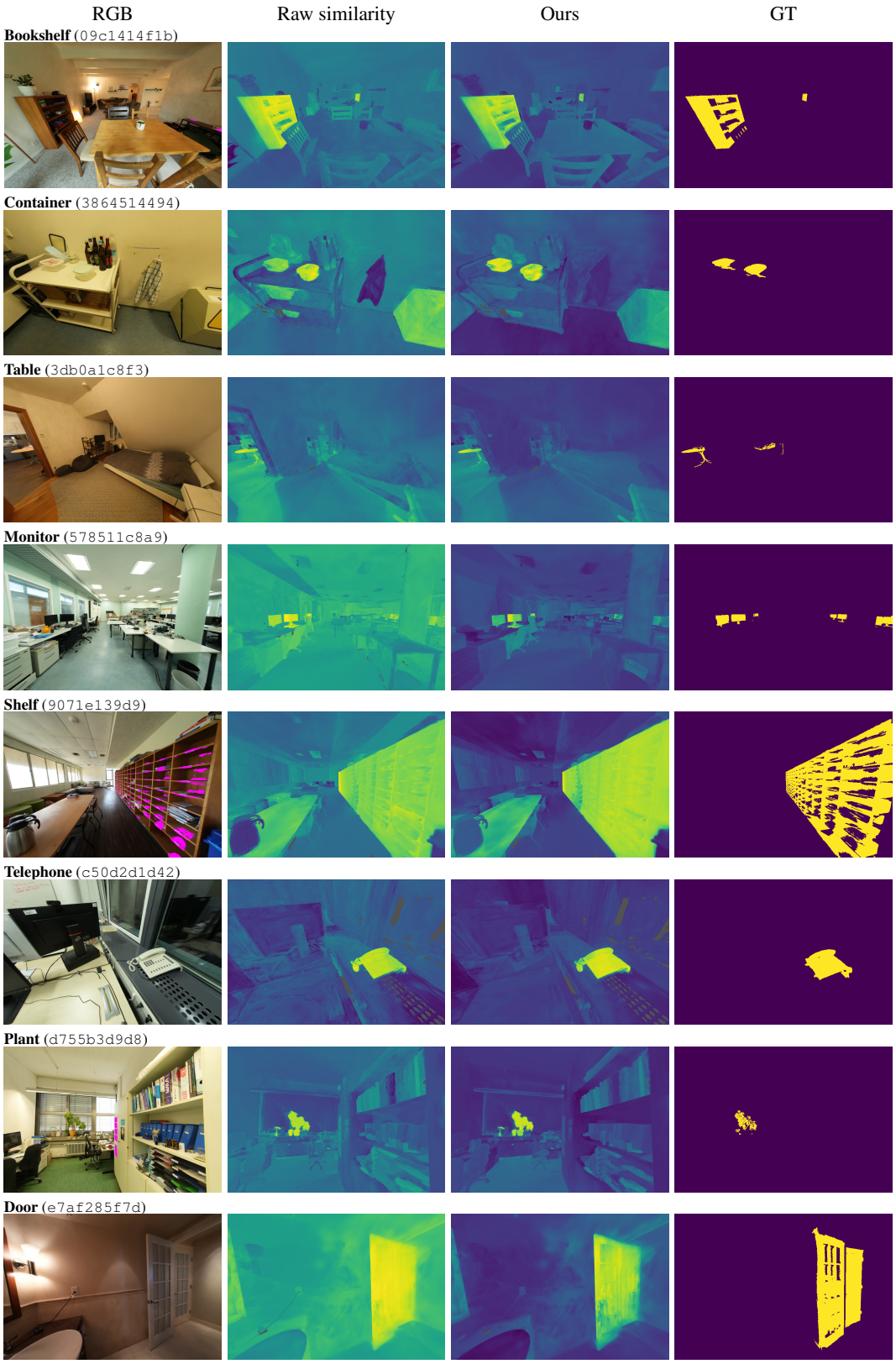


Figure 4: **Qualitative query maps on eight ScanNet++ scenes.** Each row shows one query. In mesh-derived binary GT, yellow denotes the query and purple denotes zero. Raw cosine and fully calibrated score maps are independently min–max normalized (purple: low; yellow: high). Note how the calibrated scores increase contrast between foreground and background.

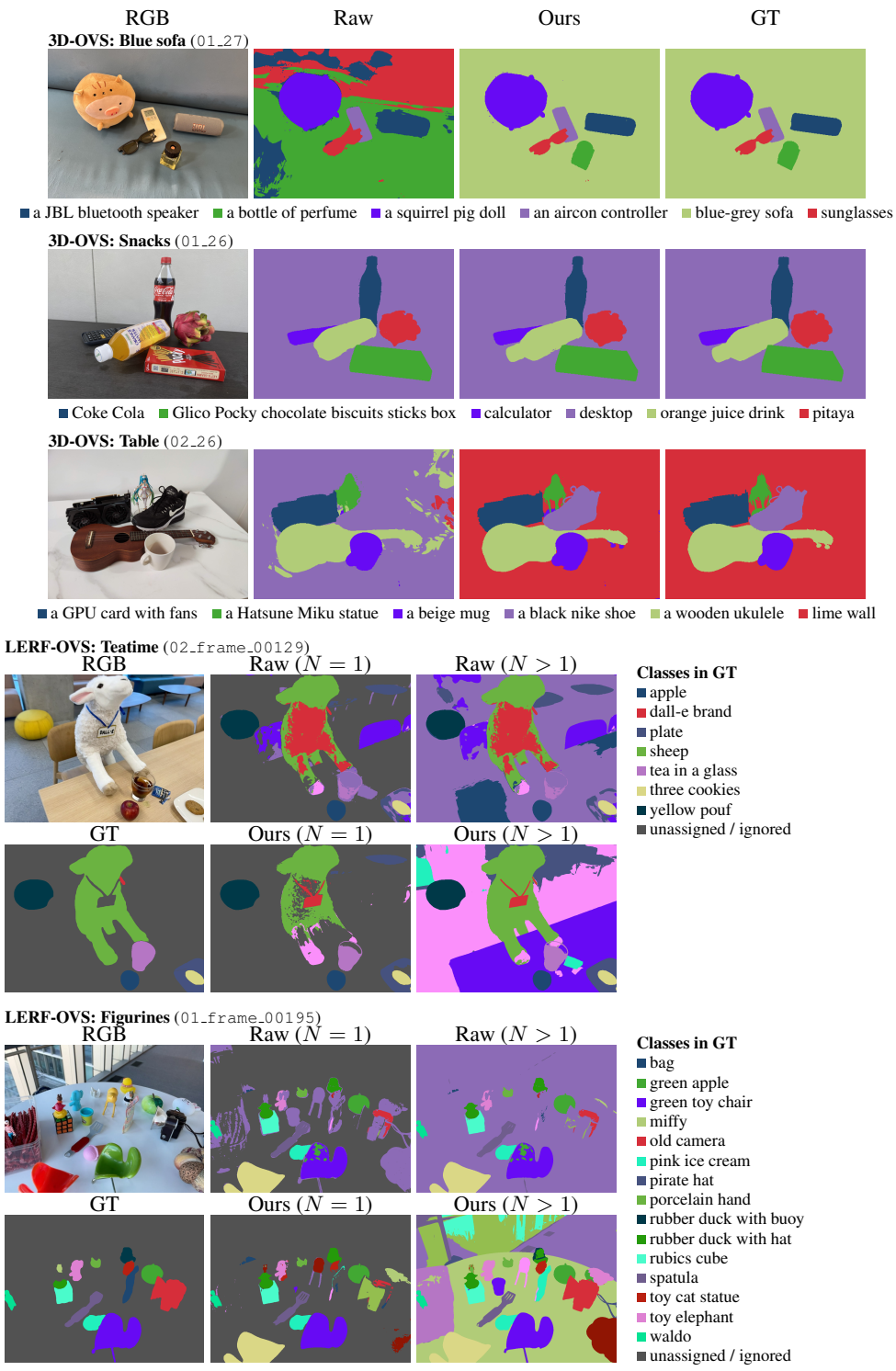


Figure 5: **Qualitative multiclass segmentation on 3D-OVS and LERF-OVS.** Each 3D-OVS row or LERF-OVS row pair shows one annotated view. For $N > 1$, argmax assigns every supported pixel to a queried class, precluding background, merged $N = 1$ masks instead retain background where all queries are rejected.

A Additional Results

Qualitative results. Figure 4 shows illustrative SFS query maps from eight ScanNet++ scenes, with one query per scene. These examples use the present classes from the 100-class vocabulary and calibration parameters $\lambda = 0.45$ and $\alpha = 0.6$. The raw and fully calibrated maps are independently min–max normalized for visualization, so they compare spatial response patterns rather than absolute score magnitudes.

More Segmentation results. We report detailed results of all our datasets in Table 5. The table shows mIoU, precision and recall for all dataset, including the smaller vocabulary versions of ScanNet (10-class) and ScanNet++ (20-class). The LERF-OVS and 3D-OVS numbers use rendered 2D evaluation following their respective protocols.

Single-query component ablations. Table 4 reports the $N = 1$ SFS component ablations with independently optimized thresholds.

Table 4: **Single-query component ablations.** We report ScanNet and ScanNet++ using SFS in the single-query setting. Each ablations threshold is fit per method and dataset using the held-out calbiration scenes. Both Whiten and CSLS depends on Bank being present, as they are no-ops on a single query. Ablations with Bank use query vs. max-reference margin, otherwise the query score is thresholded directly.

				ScanNet			ScanNet++		
Center	Whiten	CSLS	Bank	mIoU ↑	Precision ↑	Recall ↑	mIoU ↑	Precision ↑	Recall ↑
				26.93	36.44	63.20	22.65	36.00	40.78
✓				26.40	39.05	54.18	18.60	31.22	35.32
			✓	27.47	40.41	58.59	21.71	35.96	37.87
		✓	✓	28.60	40.94	60.00	24.01	36.43	41.29
✓	✓		✓	30.71	47.82	57.31	20.56	34.53	38.75
✓	✓	✓	✓	**34.52**	**47.87**	**63.77**	**24.39**	**36.14**	**43.62**

B Reference Vocabulary Bank

We use bare phrases, without a prompt template. The same bank is used across scenes, datasets, and lifting methods. Its strings have no exact, case-insensitive match to the evaluated class names, although broad semantic overlap (e.g., "furniture") is not excluded.

Table 6: **Reference Vocabulary.**

object	stuff	surface	material
background	structure	furniture	room
clutter	corner	space	interior
scene	region	pattern	shape
volume	boundary	arrangement	composition
portion	section	expanse	form
void	span	zone	patch
field	bulk	plane	ridge
groove	seam	face	side
depth	extent	scale	spacing
texture detail	contour	profile	margin
interval			

For whitening, the mean and covariance are fitted once from the 45 words. Requested queries are then transformed by this fixed reference calibration. Adding, removing, or regrouping queries does not change the fitted statistics. For $N > 1$, bank rows are discarded before class scoring. For $N = 1$, the transformed bank scores remain available for the target-minus-maximum-bank rejection margin (Section 3.5), but the query is not included in the mean or covariance estimate.

C Held-out thresholds

For the main $N = 1$ evaluations, we use thresholds determined on held-out scenes and keep them fixed across all evaluation scenes and queries. On ScanNet, Raw uses $\tau = 0.232$ for all methods except LangSplat, which uses $\tau = 0.15$. Ours uses $\tau = -0.040$ for all methods except OpenGaussian, which uses $\tau = -0.024$, and LangSplat, which uses $\tau = -0.100$. On ScanNet++, every method uses Raw $\tau = 0.245$. Ours uses $\tau = -0.020$ for SFS, LUDVIG, and OpenGaussian, $\tau = -0.022$

Table 5: **Segmentation Results Raw vs. Calibrated across All Datasets.** Each dataset pair reports multi-query ($N > 1$), followed by single-query ($N = 1$) results.

	LERF-OVS									3D-OVS								
	mIoU ↑			Precision ↑			Recall ↑			mIoU ↑			Precision ↑			Recall ↑		
Method	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ
Multi-Query $N > 1$																		
SFS	49.09	**65.67**	+16.57	67.67	**77.11**	+9.45	61.46	**71.23**	+9.77	70.79	**89.80**	+19.01	75.94	**94.09**	+18.15	88.95	**95.58**	+6.63
NormLift	58.55	**60.25**	+1.70	**74.58**	68.42	−6.16	**68.23**	66.97	−1.26	70.90	**86.11**	+15.20	74.41	**90.49**	+16.08	88.00	**93.11**	+5.10
LUDVIG	38.05	**45.44**	+7.39	**59.77**	57.75	−2.02	49.76	**58.34**	+8.57	4.15	**5.85**	+1.70	10.50	**16.80**	+6.30	14.07	**14.59**	+0.52
Occam	57.13	**61.05**	+3.92	**71.66**	69.71	−1.95	66.00	**67.21**	+1.21	66.19	**84.21**	+18.01	73.14	**89.04**	+15.91	87.94	**93.08**	+5.14
VALA	**59.83**	59.68	−0.15	**74.89**	71.72	−3.16	**68.59**	66.46	−2.12	64.16	**79.02**	+14.86	71.88	**85.49**	+13.61	87.23	**91.53**	+4.30
OpenGaussian	33.62	**46.18**	+12.56	41.57	**51.09**	+9.51	45.79	**55.00**	+9.21	23.73	**28.81**	+5.08	36.45	**45.91**	+9.46	56.93	**59.22**	+2.29
LangSplat	1.49	**5.15**	+3.66	15.03	**21.44**	+6.41	8.88	**10.24**	+1.37	7.43	**15.19**	+7.76	26.87	**41.79**	+14.93	20.38	**32.55**	+12.17
Single-Query $N = 1$																		
SFS	45.71	**53.86**	+8.15	56.91	**67.52**	+10.60	**79.10**	76.00	−3.11	69.21	**83.59**	+14.39	78.00	**90.78**	+12.78	82.18	**92.01**	+9.83
NormLift	48.82	**53.03**	+4.21	63.21	**66.89**	+3.69	**68.16**	67.82	−0.34	68.62	**82.47**	+13.86	77.83	**88.63**	+10.79	80.08	**92.31**	+12.23
LUDVIG	31.98	**35.15**	+3.16	42.49	**53.14**	+10.65	**74.02**	62.19	−11.83	13.94	**13.96**	+0.03	16.40	**16.42**	+0.03	81.98	**83.03**	+1.05
Occam	46.09	**53.67**	+7.58	64.16	**66.38**	+2.22	68.07	**69.87**	+1.80	61.10	**75.28**	+14.17	77.61	**83.39**	+5.79	73.99	**78.15**	+4.16
VALA	21.30	**54.15**	+32.85	42.98	**71.03**	+28.05	38.65	**65.88**	+27.22	31.19	**75.01**	+43.82	54.41	**84.59**	+30.19	56.48	**78.24**	+21.76
OpenGaussian	30.97	**31.06**	+0.09	37.35	**41.75**	+4.40	**56.50**	41.13	−15.37	30.72	**34.24**	+3.52	55.69	**59.15**	+3.46	36.38	**36.41**	+0.03
LangSplat	4.76	**5.19**	+0.43	8.77	**8.89**	+0.12	**77.51**	57.28	−20.23	14.96	**19.57**	+4.61	17.32	**29.49**	+12.17	**86.29**	69.34	−16.95

	ScanNet (10-class)									ScanNet (19-class)								
	mIoU ↑			Precision ↑			Recall ↑			mIoU ↑			Precision ↑			Recall ↑		
Method	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ
Multi-Query $N > 1$																		
SFS	38.25	**48.91**	+10.65	64.23	**70.97**	+6.74	55.70	**65.14**	+9.45	26.87	**39.37**	+12.50	56.38	**60.67**	+4.30	41.45	**51.18**	+9.72
NormLift	43.89	**50.18**	+6.29	63.41	**71.02**	+7.61	63.70	**66.14**	+2.45	31.40	**39.55**	+8.14	55.10	**60.68**	+5.58	48.57	**52.64**	+4.07
LUDVIG	40.70	**49.48**	+8.78	64.25	**71.97**	+7.71	59.01	**65.10**	+6.09	29.32	**39.83**	+10.51	56.48	**60.39**	+3.91	44.78	**52.81**	+8.03
Occam	34.74	**45.43**	+10.69	57.27	**67.34**	+10.07	53.68	**61.50**	+7.82	23.87	**32.11**	+8.24	50.09	**55.54**	+5.45	39.06	**45.25**	+6.19
VALA	38.94	**43.73**	+4.79	59.77	**66.78**	+7.01	58.16	**59.42**	+1.26	27.81	**36.16**	+8.35	53.13	**57.98**	+4.85	43.05	**48.98**	+5.94
OpenGaussian	33.87	**41.79**	+7.91	54.21	**66.08**	+11.87	52.33	**57.53**	+5.19	24.34	**32.09**	+7.76	46.80	**53.16**	+6.36	39.68	**43.76**	+4.08
LangSplat	5.97	**7.17**	+1.20	**30.72**	29.33	−1.39	**20.82**	18.44	−2.38	3.04	**4.00**	+0.96	**19.96**	18.55	−1.41	**11.34**	8.06	−3.28
Single-Query $N = 1$																		
SFS	30.47	**39.91**	+9.44	43.91	**55.40**	+11.50	62.12	**65.82**	+3.70	26.66	**34.52**	+7.85	39.77	**47.87**	+8.10	56.52	**63.77**	+7.25
NormLift	34.56	**42.30**	+7.74	45.81	**57.77**	+11.95	66.57	**67.20**	+0.64	30.56	**36.64**	+6.08	40.61	**49.55**	+8.94	64.74	**66.94**	+2.20
LUDVIG	31.70	**41.10**	+9.39	44.38	**56.77**	+12.39	62.95	**65.95**	+3.00	28.46	**35.76**	+7.30	40.39	**49.11**	+8.73	59.53	**64.45**	+4.91
Occam	27.66	**35.41**	+7.75	43.32	**52.78**	+9.46	50.44	**57.58**	+7.14	22.96	**31.01**	+8.05	37.33	**46.30**	+8.97	44.56	**54.45**	+9.88
VALA	30.26	**37.59**	+7.33	45.74	**55.01**	+9.27	53.03	**60.58**	+7.55	26.60	**32.85**	+6.26	40.60	**47.79**	+7.19	48.62	**57.65**	+9.03
OpenGaussian	26.84	**35.53**	+8.69	40.92	**57.90**	+16.98	51.75	**52.98**	+1.23	22.95	**30.81**	+7.86	36.56	**49.63**	+13.07	45.94	**53.48**	+7.54
LangSplat	15.02	**15.31**	+0.29	15.07	**16.46**	+1.39	**97.86**	86.32	−11.54	11.37	**11.62**	+0.25	11.41	**12.71**	+1.30	**97.91**	85.59	−12.32

	ScanNet++ (20-class)									ScanNet++ (100-class)								
	mIoU ↑			Precision ↑			Recall ↑			mIoU ↑			Precision ↑			Recall ↑		
Method	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ	Raw	Ours	Δ
Multi-Query $N > 1$																		
SFS	27.70	**33.31**	+5.61	45.81	**52.83**	+7.02	43.77	**44.79**	+1.02	19.92	**23.74**	+3.81	39.70	**41.99**	+2.28	28.23	**30.83**	+2.60
NormLift	28.89	**34.03**	+5.14	46.20	**52.76**	+6.56	44.76	**45.19**	+0.43	20.60	**24.76**	+4.16	39.19	**42.05**	+2.86	29.70	**32.22**	+2.52
LUDVIG	28.39	**34.10**	+5.71	47.25	**53.87**	+6.62	44.42	**45.75**	+1.33	20.31	**23.96**	+3.66	40.81	**41.24**	+0.43	28.54	**30.62**	+2.07
Occam	27.62	**33.30**	+5.68	44.33	**51.57**	+7.24	43.85	**45.32**	+1.46	20.18	**23.16**	+2.98	39.07	**40.35**	+1.28	29.34	**30.09**	+0.75
VALA	26.93	**31.94**	+5.01	44.25	**49.80**	+5.55	43.01	**43.93**	+0.92	19.28	**22.90**	+3.61	38.69	**40.75**	+2.06	28.27	**30.18**	+1.92
OpenGaussian	24.53	**28.36**	+3.83	37.53	**44.29**	+6.76	40.29	**40.85**	+0.56	15.25	**17.72**	+2.47	24.18	**29.95**	+5.76	23.25	**24.49**	+1.23
LangSplat	9.53	**11.59**	+2.06	20.60	**26.80**	+6.21	**19.31**	18.84	−0.47	4.50	**6.08**	+1.58	11.77	**15.80**	+4.04	8.15	**9.29**	+1.13
Single-Query $N = 1$																		
SFS	26.64	**29.92**	+3.28	37.94	**46.51**	+8.57	**54.36**	48.55	−5.81	22.55	**24.39**	+1.84	34.64	**36.14**	+1.51	42.56	**43.62**	+1.06
NormLift	28.20	**31.37**	+3.17	39.76	**45.65**	+5.89	**55.15**	51.62	−3.53	23.89	**25.54**	+1.65	35.33	**35.39**	+0.07	44.01	**46.87**	+2.87
LUDVIG	27.13	**30.90**	+3.77	37.29	**46.84**	+9.55	**56.99**	50.11	−6.89	23.11	**25.27**	+2.16	34.85	**36.83**	+1.98	44.67	**45.30**	+0.63
Occam	26.81	**30.06**	+3.25	38.99	**46.34**	+7.35	**52.98**	49.02	−3.96	22.97	**24.24**	+1.27	35.08	**35.21**	+0.13	42.66	**44.25**	+1.59
VALA	25.45	**27.42**	+1.97	39.02	**45.85**	+6.83	**48.53**	44.32	−4.22	21.99	**22.67**	+0.68	35.53	**35.61**	+0.07	39.08	**40.08**	+1.00
OpenGaussian	24.13	**27.99**	+3.86	34.16	**41.88**	+7.72	**50.56**	50.25	−0.31	18.91	**22.08**	+3.17	26.33	**30.59**	+4.26	36.94	**43.01**	+6.07
LangSplat	9.27	**11.96**	+2.69	16.59	**18.79**	+2.20	22.81	**31.57**	+8.76	5.92	**8.05**	+2.13	11.77	**12.83**	+1.07	15.65	**24.06**	+8.41

for NormLift, $\tau = -0.017$ for Occam, $\tau = -0.008$ for VALA, and $\tau = -0.040$ for LangSplat. Raw cosine scores and reference-only relevance margins have separate thresholds. We apply these fixed values directly, without refitting thresholds or imposing a recall constraint during this evaluation.

D SENSITIVITY OF λ AND α

Table 7: **Parameter sensitivity.** SFS multiclass mIoU (%) on ScanNet (19 classes) / ScanNet++ (100 classes) is shown on the top / bottom of each cell. The bold cell marks the (λ, α) chosen by held-out calibration.

$\lambda \backslash \alpha$	0.30	0.35	0.40	0.45	0.50
0.20	40.06 23.35	39.95 23.34	39.91 23.36	39.88 23.45	39.86 23.41
0.25	39.84 23.51	39.73 23.44	39.69 23.47	39.50 23.77	39.41 23.74
0.30	39.47 23.55	39.40 23.71	**39.37** **23.74**	39.35 23.70	39.38 23.69
0.35	39.16 23.48	39.14 23.41	39.12 23.49	38.88 23.50	38.87 23.51
0.40	39.38 23.80	38.74 23.64	38.71 23.64	38.68 23.71	38.69 23.70

Table 7 reports SFS multiclass sensitivity of λ and α by evaluating neighboring combinations of our selection. We selected $(\lambda, \alpha) = (0.30, 0.40)$ by averaging multiclass and single-query mIoU using held-out scenes. The table displays the sensitivity of this choice on our test sets with ScanNet on top and ScanNet++ on the bottom. The bold cell marks the selected pair, not the maximum of the displayed multiclass scores. The selection based on the calibration scenes did not find the highest scoring combination, but generally λ and α are robust to slight variations. Our LERF-OVS and OVS-3D evaluations use the same parameters.

E EVALUATION POPULATION AND CORRESPONDENCE PROTOCOLS

Gaussian-based semantic reconstructions differ both in their geometric parameterization and in the rules used to transfer primitive predictions to annotated points. For example, the official OpenGaussian ScanNet configuration (Wu et al., 2024) fixes the initial point locations and disables densification, permitting evaluation on aligned point rows. Its evaluator also ignores GT labels at rows whose Gaussian opacity is below 0.1, which can be a substantial amount in some cases. Dr. Splat (Kim et al., 2025) instead supports unconstrained Gaussian geometry and describes a Gaussian-level evaluation with Mahalanobis-based pseudo-label assignment and significance weighting.

Some prior evaluation protocols discard GT points when their associated predictions are removed by opacity or feature-validity filtering. This makes the evaluated population method-dependent and in many cases drop a substantial amount of points (see Table 8). Our evaluation instead retains every active GT point and assigns it to an eligible Gaussian using Mahalanobis distance before evaluating the semantic prediction. Let $\mathcal{E}_m$ contain the exported Gaussians of method m with opacity at least 0.1 and a valid feature under the method's existing feature-validity rule. For every active GT point $\boldsymbol{x}$, we assign

$$j_m^*(\boldsymbol{x}) = \arg\min_{j \in \mathcal{E}_m} (\boldsymbol{x} - \boldsymbol{\mu}_{mj})^\top \boldsymbol{\Sigma}_{mj}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{mj}), \qquad \widehat{y}_m(\boldsymbol{x}) = \widehat{y}_{m, j_m^*(\boldsymbol{x})}. \tag{8}$$

In this geometric equation, $\boldsymbol{x}, \boldsymbol{\mu}_{mj} \in \mathbb{R}^3$ and $\boldsymbol{\Sigma}_{mj} \in \mathbb{R}^{3\times 3}$. These geometric covariances are distinct from the text covariance in Section 3.2. Here $\boldsymbol{\Sigma}_{mj}$ is constructed from the activated Gaussian scales and rotations. Opacity is used for candidate eligibility, not as an additional factor in this distance. Neither GT class labels nor predicted semantic classes enter the association rule. The same correspondences are used for raw and calibrated predictions. Multiple GT points may select the same Gaussian.

LangSplat field adaptation. For LangSplat (Qin et al., 2024), we use the decoded 512-dimensional features from its level-3 language field as a fixed field for both Raw and Ours. The original rendered-image evaluator forms relevance maps at three levels, applies spatial smoothing, and selects a level for each query and image using the highest peak relevance. We do not reproduce that query-dependent level selection or spatial-smoothing stage as it interferes with out calibrations. The fixed field provides one representation for multiclass refinement and single-query thresholding without introducing an additional level-selection rule. The reported LangSplat results therefore evaluate this fixed-field adaptation rather than the complete original rendered-image pipeline.

Table 8: **Discarded prediction statistics.** Low opacity, invalid feature, their intersection (Both), and their union (Either) characterize ineligibility as percentages of primitives. GT dropped and GT unsupported report percentages of GT points omitted from evaluation or left without a usable prediction under prior evaluation rules. In contrast, our evaluation retains every active GT point and assigns it to an eligible Gaussian.

Dataset	Method	Low opacity	Invalid feature	Both	Either	GT dropped	GT unsupported
ScanNet	SFS	26.33	21.65	19.98	28.00	0.00	1.94
ScanNet	NormLift	26.33	21.67	19.55	28.46	0.00	2.14
ScanNet	LUDVIG	26.33	19.30	17.61	28.02	0.00	1.98
ScanNet	Occam	28.38	0.00	0.00	28.38	0.00	0.00
ScanNet	VALA	0.23	0.00	0.00	0.23	0.00	0.00
ScanNet	LangSplat	26.33	0.00	0.00	26.33	0.00	0.00
ScanNet++	SFS	30.95	10.12	10.00	31.07	28.10	28.10
ScanNet++	NormLift	30.95	6.55	6.36	31.14	28.12	28.12
ScanNet++	LUDVIG	9.50	0.03	0.00	9.53	12.46	12.46
ScanNet++	Occam	27.76	0.00	0.00	27.76	25.73	25.73
ScanNet++	VALA	0.00	0.00	0.00	0.00	0.00	0.00
ScanNet++	OpenGaussian	42.59	22.58	12.77	52.41	34.44	34.44
ScanNet++	LangSplat	30.95	0.00	0.00	30.95	28.00	28.00