GENERALIZED FEW-SHOT 3D SEGMENTATION AND MORE: SUPPORT-FREE NOVEL-CLASS DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing generalized few-shot 3D semantic segmentation methods typically assume a fixed set of novel classes and rely on noisy pointwise pseudo-labels and limited support prototypes, making missed novel-class instances difficult to recover in both spatial structure and semantics. To address this issue, we focus on novel-instance recovery directly from the query scene and propose Open-World Posterior Completion (OW-PC), a nonparametric framework built upon distance-dependent Chinese Restaurant Process (ddCRP) for structure-aware novel-class discovery. Specifically, Dual-View Debiasing (DVD) integrates complementary pretrained semantic knowledge and episodic support to extract reliable novel-class evidence, while Counterfactual Base Guard (CBG) preserves confident base-class predictions and exposes ambiguous regions for further exploration. OW-PC then organizes sparse novel evidence into coherent instance structures through spatial and semantic consistency, enabling missed novel instances to be recovered both with and without available support. Unlike conventional support-conditioned refinement, its nonparametric formulation does not require the number of novel classes to be fixed in advance, allowing potential novel structures to be actively discovered from the query scene. Experiments on ScanNet, ScanNet++, and ScanNet200 demonstrate consistent improvements in novel-class recall and harmonic mean while preserving base-class segmentation accuracy.

1 INTRODUCTION

3D semantic segmentation (Halperin & Eisl, 2025) is a fundamental capability for embodied AI (Wong et al., 2025), supporting applications such as autonomous navigation (Abou Haidar et al., 2024), robotic manipulation (Li et al., 2024), and augmented reality (Li et al., 2025). Although fully supervised methods have achieved strong performance in closed-vocabulary settings, real-world environments continually introduce categories beyond the training vocabulary, making exhaustive annotation impractical. Few-shot 3D semantic segmentation addresses this challenge by adapting to novel classes from only a handful of labeled support examples. Its generalized setting is more demanding: base classes remain present together with novel classes in unseen query scenes, requiring the model to recognize both simultaneously (Xu et al., 2023; An et al., 2025a). The central challenge is therefore not only to recognize novel classes from limited supervision, but also to recover their spatial extent without incorrectly absorbing known base regions, thereby balancing novel-class discovery and base-class preservation.

Existing methods mainly rely on support prototypes, VLM-derived semantic priors, and pseudo-label refinement to recognize novel classes (Zhao et al., 2021; He et al., 2023; An et al., 2024; 2025b;a; Peng et al., 2023; Nguyen et al., 2024). Support-prototype methods construct class representations from a few labeled examples, but limited support can hardly capture substantial intra-class variation between support and query instances. VLM-enhanced methods introduce pretrained semantic knowledge through pointwise predictions or pseudo-labels, broadening the recognizable semantic space but remaining susceptible to occlusion, view-projection errors, and bias toward familiar base categories. Subsequent prototype refinement and pseudo-label propagation can expand already identified novel regions, yet their search space is still largely determined by the initial predictions and support-derived candidates. Once a true novel region is incorrectly assigned to a base class or background and excluded from the reliable novel candidates, limited support evidence may be insufficient to recover it in later stages. As illustrated in Fig. 1(b), the novel stool is absorbed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

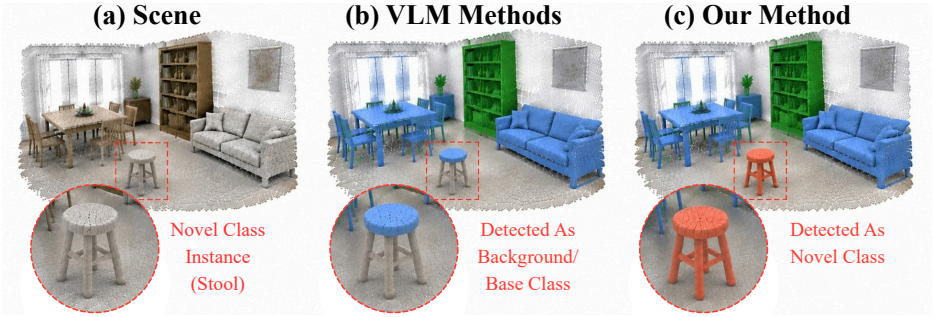


Figure 1: Motivation of novel-instance recovery on a query scene. Blue and green denote base classes; orange-red denotes a novel class with only a few labeled support examples. (a) The scene contains a novel-class instance, a stool, that is absent from the base vocabulary. (b) Existing VLM-based methods (An et al., 2025a; Peng et al., 2023; Nguyen et al., 2024), affected by base-class bias, occlusion, and view projection, predict the stool as background or a base class, causing the novel instance to be missed. (c) Our method recovers the missed instance directly from the query scene and correctly identifies it as a novel-class instance.

into background or base-class predictions, causing not only incorrect pointwise labels but the loss of an entire novel-instance structure from subsequent discovery. Moreover, existing methods typically assume that the novel vocabulary and the number of novel classes are fixed by the available support, whereas real-world query scenes may contain an unknown and varying number of unsupported categories. When support is unavailable, incomplete, or unrepresentative, this assumption further prevents the model from discovering potential novel structures outside the predefined category space. These limitations reveal a common bottleneck: existing methods primarily refine novel evidence within a support-defined category space, rather than actively discovering and structurally recovering missed novel classes from the query scene itself.

To overcome this bottleneck, we shift from pointwise pseudo-label refinement to structured novel-instance recovery in the query scene and propose an unsupervised framework built upon the distance-dependent Chinese restaurant process (ddCRP). Our key idea is to retain reliable semantic evidence while reopening ambiguous regions for structural exploration, allowing novel instances rejected by initial predictions to re-enter the discovery process. Specifically, Dual-View Debiasing (DVD) integrates complementary pretrained semantic knowledge and episodic support to extract reliable novel-class evidence, while Counterfactual Base Guard (CBG) preserves confident base predictions and exposes potentially unknown regions for further exploration. Building on these reliable cues, our Open-World Posterior Completion (OW-PC) exploits spatial and semantic consistency to organize sparse novel evidence into coherent instance structures, thereby recovering missed novel instances rather than merely propagating existing pseudo-labels. As shown in Fig. 1(c), our method recovers the complete stool structure missed by VLM-based predictions while preserving the surrounding base-class regions. Furthermore, the nonparametric nature of ddCRP allows novel structures to emerge without fixing their number in advance. When support is available, recovered structures can be grounded to supported novel semantics; when support is absent or incomplete, scene-internal structural and feature consistency still enables category-agnostic unknown instances to be discovered. In this way, our framework reduces the dependence on noisy pointwise pseudo-labels, limited support prototypes, and a predefined novel category space, extending generalized few-shot 3D segmentation toward open-world, structure-aware novel-class discovery.

Our main contributions are summarized as follows:

- We propose Dual-View Debiasing (DVD) and Counterfactual Base Guard (CBG) to establish reliable novel evidence and a safe exploration space, respectively, enabling potential novel regions to be explored while preserving confident base-class predictions.
- We introduce Open-World Posterior Completion (OW-PC), which organizes sparse novel evidence into coherent instance structures under a nonparametric formulation. It supports both support-guided novel-class recognition and support-free unknown-instance discovery without requiring the number of novel categories to be fixed in advance.

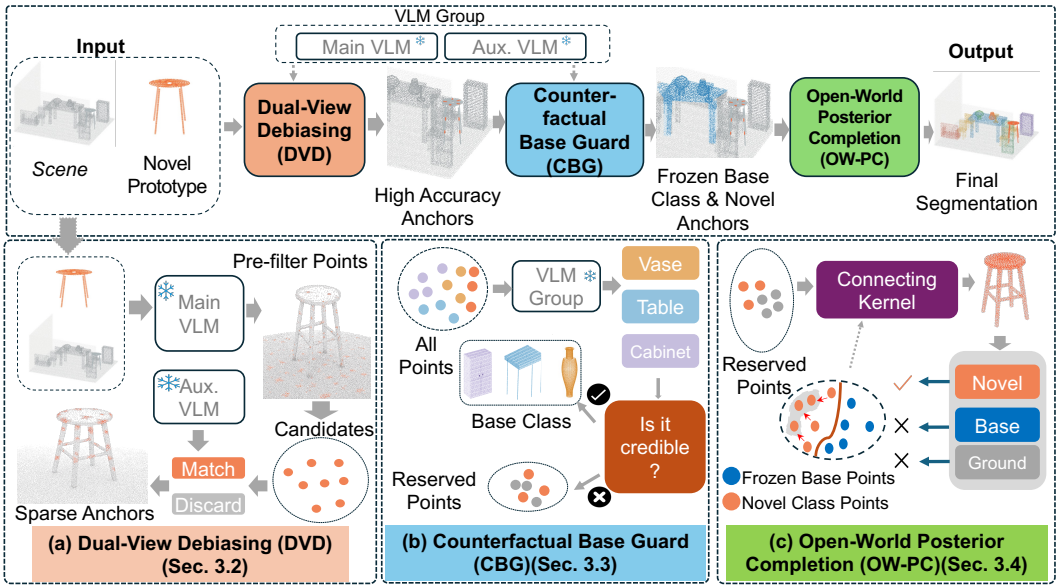


Figure 2: Overview of the proposed framework. Given a novel-class prototype and a query scene, a main and an auxiliary VLM provide complementary semantic evidence. (a) Dual-View Debiasing screens main-VLM proposals and has the auxiliary VLM adjudicate them, yielding high-accuracy novel-class anchors. (b) Counterfactual Base Guard keeps credible base predictions and reserves the remaining points for structural inference. (c) Open-World Posterior Completion groups anchors and reserved points with a ddCRP connecting kernel, grows novel instances, and applies a Bayes-factor judge to output the final novel/base/ground segmentation.

- Extensive experiments on ScanNet, ScanNet++, and ScanNet200 demonstrate consistent improvements in novel-class recall and harmonic mean while preserving base-class segmentation performance.

2 Related Work

Few-shot and generalized few-shot 3D segmentation. Early few-shot 3D methods transfer base-class representations to novel point-cloud classes through prototype matching, attention-based aggregation, or class-specific correlations (Zhao et al., 2021; He et al., 2023; Zhang et al., 2023; An et al., 2024; Wang et al., 2026a). Generalized few-shot segmentation advances this setting by requiring base and novel classes to be segmented jointly at test time (Xu et al., 2023; An et al., 2025a). Recent multimodal variants further introduce VLM semantic priors for support–query adaptation (An et al., 2025b;a; Wang et al., 2026b); related open-vocabulary 3D systems lift multi-view language-aligned features or masks to point clouds (Radford et al., 2021; Peng et al., 2023; Lee et al., 2025; Tao et al., 2026) and aggregate them into object proposals or open-set maps (Takmaz et al., 2023; Nguyen et al., 2024; Jatavallabhula et al., 2023; Lee et al., 2026). Support representation is central throughout this development. Feature-hallucination methods augment limited support sets (Pahde et al., 2019), distribution-calibration methods transfer statistics from data-rich base classes to data-scarce novel classes (Yang et al., 2021), and probabilistic prototypes explicitly represent support uncertainty (Liu et al., 2025). These designs strengthen semantic recognition from limited supervision, but novel-region recovery still generally depends on dense VLM candidates or pointwise prototype matching. We instead retain observed support modes as finite, uncertainty-tempered semantic evidence and use them to explain candidate components recovered from the unlabeled query scene.

Bayesian and open-world structural discovery. Dirichlet processes and their Chinese restaurant process representation induce partitions whose number of occupied components can grow with the observations (Ferguson, 1973; Antoniak, 1974). The distance-dependent CRP (ddCRP) replaces

exchangeable seating with observation-dependent links, inducing nonexchangeable partitions from spatial or feature proximity (Blei & Frazier, 2011); spatial variants use these local cues for image segmentation (Ghosh et al., 2011). In few-shot learning, transductive methods exploit joint support–query structure but typically propagate labels over a fixed propagation topology or enrich a fixed set of prototypes (Garcia & Bruna, 2018; Liu et al., 2019; Thengane et al., 2026). In parallel, unsupervised 3D segmentation uses view-consistent semantic distillation to expose coherent scene regions without class annotations (Song et al., 2026; Zhang et al., 2026). OW-PC combines adaptive local partitioning with a support-availability conditional component posterior: DVD anchors constrain feasible named explanations, CBG supplies a protected domain and base counterfactuals, and marginal evidence evaluates every recovered component against named, base, background, and unknown alternatives.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We follow the generalized few-shot 3D semantic segmentation protocol (Xu et al., 2023; An et al., 2025a). Let $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ denote base classes (with full training annotations) and novel classes (unseen during training), respectively. For each novel class $c \in \mathcal{C}_{\text{novel}}$, we are given $K \in \{1, 3, 5\}$ support point clouds $\{S_c^{(k)}\}_{k=1}^{K}$ with ground-truth labels. At test time, given a query scene Q, the model assigns a per-point label $y_i \in \mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}} \cup \{\text{unlabeled}\}$ for each point $i \in Q$. A frozen primary VLM provides a language-aligned point feature $f_i \in \mathbb{R}^{d_f}$ and primary class distribution p_i^{pri}, while a separately pretrained auxiliary VLM provides an auxiliary distribution p_i^{aux} (Choy et al., 2019; Peng et al., 2023). OW-PC uses the same frozen query feature together with voxel coordinates $x_i \in \mathbb{R}^3$.

To recover support-defined novel classes from the unlabeled query structure, the stages in Figure 2 have complementary roles: (1) **Dual-View Debiasing (DVD)** (Sec. 3.2) uses the primary VLM and episode-specific support to generate and screen class-conditioned candidates, then uses the auxiliary VLM for second-pass semantic adjudication. A fused posterior rejects inconsistent candidates and assigns the final class of each retained anchor, yielding a sparse novel seed set $\mathcal{S} = \{(i, c) : i \in Q, c \in \mathcal{C}_{\text{novel}}\}$. (2) **Counterfactual Base Guard (CBG)** (Sec. 3.3) protects calibrated high-confidence base elements and exposes only ambiguous base predictions, together with unlabeled elements, to structural re-evaluation. (3) **Open-World Posterior Completion (OW-PC)** (Sec. 3.4) organizes sparse DVD anchors and compatible CBG-exposed elements into instance structures, jointly inferring their partition and support-availability conditional component state. The support set instantiates named semantic hypotheses; when it is absent or partial, a category-agnostic unknown hypothesis is added without changing structural inference. Anchored components expand into compatible neighborhoods, while sufficiently decisive unanchored components recover unknown instances; a ddCRP prior supplies the local link potential.

3.2 DUAL-VIEW DEBIASING: ASYMMETRIC VARIATIONAL ANCHOR DISTILLATION

DVD addresses unreliable, base-biased novel predictions from a single VLM, which can contaminate later instance recovery. It uses support-conditioned primary proposals and independent auxiliary verification to retain high-precision, class-aware anchors for OW-PC, formulating candidate correctness through variational inference.

Primary-VLM candidate screening. Let $p_i^{\text{pri}}(c)$ and f_i denote the primary VLM probability and language-aligned feature of query voxel i. For support instance k, let $\bar{f}_c^{(k)}$ be the mean primary VLM feature of its labeled class-c points. We define episode-specific compatibility by

$$s_i^c = \exp\left(-\frac{\min_k \|f_i - \bar{f}_c^{(k)}\|_2^2}{2\sigma_s^2}\right). \tag{1}$$

Here $p_i^{\text{pri}}(c)$ expresses category-level compatibility, whereas s_i^c expresses resemblance to the novel instances in the current episode. Let $z_i^c \in \{0, 1\}$ indicate whether the class-c proposal at voxel i is a valid anchor, and define $\text{logit}_\delta(q) = \log \frac{q+\delta}{1-q+\delta}$. The primary view and episodic support form its

initial proposal log-odds

$$L_i^{c,0} = b_0 + \lambda_{\text{pri}} \operatorname{logit}_\delta(p_i^{\text{pri}}(c)) + \lambda_S \operatorname{logit}_\delta(s_i^c), \tag{2}$$

inducing an initial proposal distribution $q_i^0(z_i^c = 1) = \operatorname{sigmoid}(L_i^{c,0})$. Thus the support prototype conditions proposals within the primary embedding instead of acting as another model vote.

Auxiliary semantic adjudication as a likelihood ratio. Let $u_i^c = \operatorname{rel}(\hat{y}_i^{\text{aux}}, c)$ summarize the semantic relation between the auxiliary prediction and candidate class c as one of four discrete states: match, another novel class, a base class, or other. On held-out base-class episodes, where subsets of base classes are treated as pseudo-novel tasks, we estimate with smoothing

$$\psi_u^+ = P(u_i^c = u \mid z_i^c = 1), \qquad \psi_u^- = P(u_i^c = u \mid z_i^c = 0). \tag{3}$$

The auxiliary observation contributes the cross-view likelihood ratio

$$\Lambda_i^{\text{aux},c} = \frac{\psi_{u_i^c}^+ + \delta}{\psi_{u_i^c}^- + \delta}. \tag{4}$$

Unlike a binary agreement test, $\Lambda_i^{\text{aux},c} > 1$ reinforces a proposal when the observed auxiliary relation is characteristic of correct anchors, while $\Lambda_i^{\text{aux},c} < 1$ suppresses it when characteristic of false-positive noise. Evaluating this evidence for every candidate class lets the auxiliary VLM serve as a second-pass semantic adjudicator: it rejects a primary candidate that is inconsistent with independent evidence and favors the class whose relation is most characteristic of a correct anchor.

Variational reliability distillation. We seek an optimal belief distribution $q_i(z_i^c)$ for anchor selection by minimizing a KL-regularized cross-view verification energy (see Appendix A.1 for the derivation). Its closed-form optimum gives the anchor log-odds

$$R_i^c = \log \frac{q_i^*(z_i^c = 1)}{q_i^*(z_i^c = 0)} = L_i^{c,0} + \lambda_{\text{aux}} \log \Lambda_i^{\text{aux},c}, \tag{5}$$

where $\lambda_{\text{aux}} \geq 0$ is a calibrated temperature discounting correlated semantics between the two pretrained representations. The resulting posterior provides both anchor reliability and a final class ranking. The high-precision anchor set is extracted by thresholding:

$$\hat{c}_i = \arg\max_{c \in \mathcal{C}_{\text{novel}}} R_i^c, \qquad \mathcal{S}_c = \{i : \hat{c}_i = c,\ q_i^*(z_i^c = 1) > \tau_R\}. \tag{6}$$

Each retained anchor receives its class identity from the fused posterior ranking $\hat{c}_i$, rather than from the primary VLM alone. DVD deliberately retains only high-reliability anchors; the remaining structure is handled by CBG and OW-PC, so anchor precision, base stability, and scene-level recall are controlled by separate variables rather than traded through one dense pseudo-label threshold.

3.3 Counterfactual Base Guard (CBG): Selective Base Protection

CBG balances base preservation and correction. It fixes only cross-view-consistent, high-margin base predictions, while leaving uncertain base and unlabeled elements in an exposed domain for structural re-evaluation, and it provides the base counterfactual used by OW-PC.

For $\ell \in \mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}}$, we combine the two pretrained views with a calibrated evidence temperature,

$$G_i(\ell) = \log\big(p_i^{\text{pri}}(\ell) + \delta\big) + \lambda_B \log\big(p_i^{\text{aux}}(\ell) + \delta\big). \tag{7}$$

Let $\hat{y}_i^{\text{pri}} = \arg\max_\ell p_i^{\text{pri}}(\ell)$ and $\hat{y}_i^{\text{aux}} = \arg\max_\ell p_i^{\text{aux}}(\ell)$. We define $\hat{b}_i = \arg\max_{b \in \mathcal{C}_{\text{base}}} G_i(b)$ and let

$$g_i = G_i(\hat{b}_i) - \max_{\ell \neq \hat{b}_i} G_i(\ell) \tag{8}$$

be its margin against all base and novel alternatives. The protected set is

$$\mathcal{P}_B = \{i : \hat{y}_i^{\text{pri}} = \hat{y}_i^{\text{aux}} = \hat{b}_i,\ g_i > \tau_b\}, \qquad \mathcal{E} = Q \setminus \mathcal{P}_B. \tag{9}$$

The threshold τ_b and temperature λ_B are calibrated on held-out base episodes. Elements in $\mathcal{P}_B$ retain their base labels and are excluded from structural growth. The exposed set $\mathcal{E}$ contains unlabeled elements and ambiguous base predictions; exposure alone never changes a label.

CBG also supplies the counterfactual explanation for a component $A \subseteq \mathcal{E}$:

$$e_A^B = \max_{b \in \mathcal{C}_{\text{base}}} \frac{1}{|A|} \sum_{i \in A} G_i(b). \tag{10}$$

This asks how well the strongest coherent base class explains the same component. OW-PC can revise an ambiguous base prediction only when a novel-class explanation defeats this component-level counterfactual, rather than merely exceeding a pointwise threshold. Protected base elements are hard and cannot be absorbed, while exposed elements join a neighboring component only when the resulting partition improves the marginal OW-PC posterior and the decoded explanation passes its LBF test; the resulting base–novel, base–background, and novel–background boundaries are discussed in Appendix A.6.

3.4 Open-World Posterior Completion (OW-PC)

OW-PC completes sparse anchors within the ambiguous exposed domain, where unassigned elements may belong to missed novel instances, base classes, or background. Its ddCRP prior groups spatially and semantically compatible elements into adaptive components, and it marginalizes the competing explanations to accept only components supported by decisive posterior evidence.

Uncertainty-aware support evidence. Let $\mathcal{C}_S$ be the classes with available labeled support. Labeled support defines a named hypothesis but is too limited to identify an unrestricted feature distribution. We retain at most $M_{\max}$ observed modes by farthest-point initialization followed by spherical assignment to normalized centroids μ_{cm}, and set each mode weight to $\pi_{cm} = n_{cm}/N_c$ and its temperature from its dispersion and support count,

$$d_{cm} = \frac{1}{n_{cm}} \sum_{n: z_{cn} = m} \left(1 - f_{cn}^\top \mu_{cm}\right), \qquad T_{cm} = T_0 + \gamma_d d_{cm} + \frac{\gamma_n}{\sqrt{n_{cm}}}, \tag{11}$$

so that dispersed or weakly observed modes are discounted. For a candidate component A, its novel-class evidence is the uncertainty-tempered mixture

$$e_A^N(c) = \log \sum_{m=1}^{M_c} \pi_{cm} \exp\left(\frac{1}{|A| T_{cm}} \sum_{i \in A} f_i^\top \mu_{cm} \right), \qquad \pi_{cm} = \frac{n_{cm}}{N_c}. \tag{12}$$

This preserves multiple observed appearances while discounting unreliable modes; the support statistics $\mu_{cm}, T_{cm}, \pi_{cm}$ are never updated from query components, so novel semantics remain defined by labeled support.

Support-availability conditional posterior completion. OW-PC induces a partition $\Pi(a)$ on $\mathcal{E}$ with a local ddCRP prior,

$$\phi_i(a_i = j) \propto \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma_x^2} - \frac{\|f_i - f_j\|_2^2}{2\sigma_f^2} \right) \mathbf{1}[j \in \mathcal{N}(i)] + \alpha\, \mathbf{1}[j = i]. \tag{13}$$

For each component A, support availability determines the competing explanations and their energy:

$$\mathcal{H}(S) = \mathcal{C}_S \cup \{\mathrm{B}, \mathrm{bg}\} \cup \{\mathrm{U} : \xi_U = 1\}, \tag{14}$$

$$E_A(h; S) = \mathbf{1}[h = c \in \mathcal{C}_S]\, e_A^N(c) + \mathbf{1}[h = \mathrm{U}]\, e_A^U + \mathbf{1}[h = \mathrm{B}]\, e_A^B + \mathbf{1}[h = \mathrm{bg}]\, e_A^{\mathrm{bg}}. \tag{15}$$

Here e_A^U measures structural coherence, e_A^B is the base counterfactual, and e_A^{bg} is a calibrated background score. The switch ξ_U enables the category-agnostic unknown hypothesis U, which names no class, when support is absent or partial. Let C_A be the anchor labels in A; anchors constrain, but do not predetermine, a component's semantic explanation:

$$\mathcal{H}_A(S) = \mathbf{1}[C_A = \{c\}]\, \{c\} \cup \mathbf{1}[C_A = \emptyset]\, \mathcal{H}(S), \tag{16}$$

A component with conflicting anchor classes therefore receives zero posterior mass. The key partition posterior is

$$\Psi(A; S) = \sum_{h \in \mathcal{H}_A(S)} \pi_h \exp(E_A(h; S)), \tag{17}$$

$$p(a \mid Q, S) \propto \prod_{i \in \mathcal{E}} \phi_i(a_i) \prod_{A \in \Pi(a)} \Psi(A; S). \tag{18}$$

Marginalizing explanations during partition inference prevents premature pointwise labeling. After decoding $\hat{h}_A = \arg\max_{h \in \mathcal{H}_A(S)} \{\log \pi_h + E_A(h; S)\}$, we compare the selected explanation against all counterfactuals,

$$\mathrm{LBF}_A(h) = \log \pi_h + E_A(h; S) - \log \sum_{h' \in \mathcal{H}(S) \setminus \{h\}} \pi_{h'} \exp\big(E_A(h'; S)\big), \tag{19}$$

and accept a named or unknown component only if

$$|A| \geq n_{\min}(\hat{h}_A) \quad \text{and} \quad \mathrm{LBF}_A(\hat{h}_A) \geq \delta_{\hat{h}_A}. \tag{20}$$

Thus OW-PC may revise exposed, but never protected, base elements.

Category cardinality and support-free completion. The inference above never fixes the number of novel classes, since the ddCRP self-link creates components as needed and retains them only when accepted; deduplicating the decoded named components therefore yields the estimated novel vocabulary without a separate classifier or count-specific threshold (Appendix A.5). When no support is available ($\mathcal{C}_S = \emptyset$, $\xi_U = 1$), the hypothesis set reduces to $\mathcal{H}_{\mathrm{sf}} = \{\mathrm{U}, \mathrm{B}, \mathrm{bg}\}$ with unknown evidence $e_A^{\mathrm{U}} = \lambda_{\mathrm{coh}} \operatorname{coh}(A) - \lambda_\partial \operatorname{cost}_\partial(A)$, where $\operatorname{coh}(A)$ aggregates compatibility over internal links and $\operatorname{cost}_\partial(A)$ penalizes links crossing strong discontinuities. The same coordinate-MAP inference recovers category-agnostic unknown instances (Appendix A.4); all priors and thresholds are calibrated on held-out base episodes and then fixed.

3.5 Training Objective

The model learns from reliability-weighted anchors and accepted component targets:

$$\mathcal{L}(\theta) = -\sum_c \sum_{i \in \mathcal{S}_c} w_i^c \log p_\theta(c \mid x_i) - \beta \sum_{i \in \mathcal{Y}} v_i \log p_\theta(\hat{y}_i \mid x_i). \tag{21}$$

Here w_i^c and v_i are anchor and component confidences, and $\mathcal{Y}$ is the set of accepted component targets. Both target sets are produced by the frozen VLM pipeline and the labeled support, so no query annotation is used. Support modes stay fixed, so query features never alter the labeled semantic evidence.

Inference and calibration. Given the anchors and the exposed domain, OW-PC optimizes the partition posterior by coordinate ascent: each exposed element is linked to itself or a feasible neighbor, and the link is kept only if it improves the marginal posterior of Equation 18. All explanation priors π_h, minimum sizes $n_{\min}(h)$, log-Bayes-factor margins δ_h, temperatures, and thresholds are estimated once on held-out base episodes converted into pseudo-novel tasks and then frozen; no query label, unsupported class name, or test statistic enters calibration. When two touching objects share similar frozen features, the local kernel may over-link them, so we additionally derive a boundary-aware link potential (Appendix A.6) that augments Equation 13 with surface-orientation, feature, and calibrated-semantic disagreement.

4 Experiments

4.1 Experiments Setup

We evaluate on S3DIS (Armeni et al., 2016), ScanNet (Dai et al., 2017), ScanNet++, and ScanNet200 (Rozenberszki et al., 2022), following the GFS-VL protocol (An et al., 2025a). We report base and novel mIoU and their harmonic mean (HM). Table 1 compares our method with PIFS (Cermelli et al., 2021), attMPTI (Zhao et al., 2021), GW (Xu et al., 2023), and GFS-VL (An et al., 2025a). The base–novel partitions follow the standard GFS-VL splits (13/6, 12/45, 12/18, and 10/3 classes on ScanNet, ScanNet200, ScanNet++, and S3DIS, respectively), and we evaluate both 1-shot and 5-shot episodes with support features frozen as fixed semantic evidence. All thresholds are calibrated on held-out base episodes and then fixed before novel-class evaluation. All experiments are conducted on two NVIDIA RTX 4090 GPUs using mixed-precision (FP16) training.

Table 1: Comparison with prior methods. Best in **bold**, second-best underlined.

	ScanNet						ScanNet200					
	1-shot			5-shot			1-shot			5-shot		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
PIFS (TIP'23)	35.80	2.54	4.75	35.14	3.21	5.88	28.78	3.82	6.71	17.84	2.87	4.88
attMPTI (CVPR'21)	12.97	1.62	2.88	16.31	3.12	5.21	37.13	4.99	8.79	54.84	3.28	6.17
GW (ICCV'23)	40.06	14.78	21.55	40.18	18.58	25.39	59.28	8.30	14.55	55.23	6.47	11.56
GFS-VL (CVPR'25)	78.56	49.72	60.88	78.30	51.22	61.91	67.57	31.67	43.12	**68.48**	29.18	40.92
Ours	**78.65**	**51.18**	**62.00**	**78.37**	**52.69**	**63.02**	**68.26**	**35.05**	**46.31**	67.52	**35.85**	**46.83**

	S3DIS						ScanNet++					
	1-shot			5-shot			1-shot			5-shot		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
PIFS (TIP'23)	–	–	–	–	–	–	36.66	4.95	8.71	39.98	5.74	10.03
attMPTI (CVPR'21)	26.42	2.77	4.99	26.82	3.49	6.17	53.16	3.55	6.66	55.89	4.19	7.78
GW (ICCV'23)	42.37	11.79	18.44	42.34	8.52	14.05	46.71	6.63	11.59	51.35	11.03	18.15
GFS-VL (CVPR'25)	**59.63**	32.69	42.22	58.83	32.31	41.71	61.39	19.42	29.47	60.05	21.66	31.82
Ours	59.36	**35.06**	**43.97**	**59.27**	**32.60**	**42.06**	**64.19**	**23.19**	**34.07**	**61.27**	**28.99**	**39.35**

Table 2: Novel-class category count.

Dataset	Ground-truth novel categories	Our prediction
S3DIS	3	3.2
ScanNet	6	5.4
ScanNet200	45	34.6
ScanNet++	18	15.4

Table 3: Module ablation on ScanNet 1-shot.

DVD	CBG	OW-PC	Base	Novel	HM
–	–	–	67.31	26.26	37.78
✓	–	–	73.95	41.93	53.51
✓	✓	–	78.71	46.40	58.38
✓	–	✓	72.93	48.31	58.12
✓	✓	✓	**78.86**	**53.47**	**63.75**

4.2 Experiments Results

Table 1 reports the comparative evaluation on ScanNet, ScanNet++, ScanNet200, and S3DIS under both 1-shot and 5-shot settings; all values are percentages.

Base-class stability. Relative to GFS-VL, our method improves base mIoU in six of the eight dataset–shot settings in Table 1; on ScanNet200 5-shot and S3DIS 1-shot, it remains close to GFS-VL (67.52 versus 68.48 and 59.36 versus 59.63). CBG freezes only calibrated high-confidence base elements, so ambiguous base predictions remain open to component-level revision.

Novel-class gains and their sources. Our method outperforms GFS-VL on novel mIoU and HM in every dataset–shot setting in Table 1. The HM gains are largest on ScanNet++ (4.60 and 7.53 points for 1- and 5-shot) and ScanNet200 (3.19 and 5.91 points), and more modest on ScanNet and S3DIS (0.35–1.75 points), driven by the mechanisms analyzed in the ablation.

4.3 Novel-Class Category Count

Table 2 reports the ground-truth and predicted novel-category counts, where "Ground-truth" is the split's class count and "Our prediction" the five-run decoded average. The prediction is near-exact for small vocabularies (3.2 vs. 3 on S3DIS; 5.4 vs. 6 on ScanNet) but underestimates larger ones, by 2.6 categories on ScanNet++ (15.4 vs. 18) and 10.4 on ScanNet200 (34.6 vs. 45). The gap therefore grows with vocabulary size: the estimate is reliable for compact label sets and increasingly conservative as the vocabulary expands. The full protocol is in Appendix A.5.

4.4 Ablation Studies

Table 3 verifies the complementary benefits of the proposed modules; all values are percentages. Without any module, the backbone is strongly base-biased and recovers few novel points (Novel 26.26). DVD raises Novel to 41.93 by removing base-biased false positives and supplying reliable novel anchors rather than dense pseudo-labels, but on its own it also erodes base accuracy (Base 73.95, or 72.93 once OW-PC is added). CBG removes this erosion, restoring Base to 78.71–78.86

Table 4: Support-free instance-level recovery. Best in **bold**, second-best underlined.

Dataset	Method	Instance recall @ IoU 0.50	Mean best IoU
ScanNet200	attMPTI	12.04 (1,173 / 9,740)	24.51
	GW	10.35 (1,008 / 9,740)	19.61
	GFS-VL	26.01 (2,088 / 8,029)	35.69
	Ours	**26.96 (2,165 / 8,029)**	**38.35**
ScanNet	attMPTI	13.83 (863 / 6,240)	25.69
	GW	9.01 (562 / 6,240)	19.98
	GFS-VL	24.63 (1,537 / 6,240)	33.19
	Ours	**25.52 (1,592 / 6,240)**	**37.46**
ScanNet++	attMPTI	11.42 (364 / 3,186)	24.83
	GW	8.10 (258 / 3,186)	18.61
	GFS-VL	17.54 (559 / 3,187)	29.23
	Ours	**20.28 (646 / 3,187)**	**31.96**
S3DIS	attMPTI	18.90 (442 / 2,339)	33.62
	GW	15.56 (364 / 2,339)	26.33
	GFS-VL	21.12 (494 / 2,339)	31.34
	Ours	**73.32 (1,715 / 2,339)**	**69.07**

Table 5: Partial-prototype semantic evaluation. Best in bold, second-best underlined.

Dataset	Method	Base	Novel with support	Novel without support	All novel	All classes
ScanNet	attMPTI	4.79	0.09	0.00	0.04	3.13
	GW	3.42	0.08	0.00	0.04	2.24
	GFS-VL	77.74	59.77	7.05	33.41	63.74
	Ours	**78.99**	**66.26**	**37.02**	**51.64**	**70.36**
ScanNet++	attMPTI	1.96	0.45	0.00	0.22	0.28
	GW	3.67	0.53	0.00	0.26	0.49
	GFS-VL	61.08	21.60	4.01	12.81	32.12
	Ours	**63.02**	**29.72**	**16.91**	**23.32**	**39.20**
ScanNet200	attMPTI	0.92	0.26	0.00	0.13	0.08
	GW	2.28	0.15	0.00	0.07	0.15
	GFS-VL	67.73	29.78	7.08	18.18	28.61
	Ours	**68.85**	**34.31**	**25.54**	**29.83**	**38.04**
S3DIS	attMPTI	4.44	0.00	0.00	0.00	3.41
	GW	2.85	0.00	0.00	0.00	2.19
	GFS-VL	58.38	42.78	42.89	42.85	54.80
	Ours	**59.42**	**47.94**	**46.64**	**46.41**	**55.03**

while further lifting Novel, confirming the value of base protection. OW-PC then grows the remaining anchors and exposed elements into coherent components, and combining all three gives the best Novel and HM (53.47 and 63.75) at Base 78.86.

4.5 Support-Availability Analysis

Standard GFS evaluation assumes complete novel support and cannot test recovery without semantic supervision. We therefore add two regimes: partial support, which withholds the prototypes of a subset of novel classes, and zero support, which supplies no novel text, prototype, or semantic logit and lets OW-PC return category-agnostic unknown instances scored post-hoc.

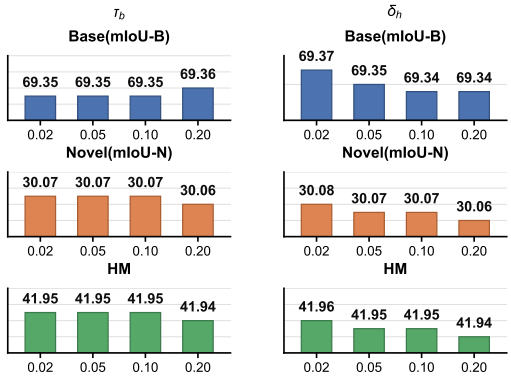


Figure 3: CBG margin τ_b (left) and ddCRP margin δ_h (right) sensitivity on ScanNet200 1-shot.

Tables 4 and 5 report matched/total recall@IoU 0.50, mean best IoU, and with/without-support mIoU. Under partial support, OW-PC improves every novel metric over GFS-VL, most on the classes whose prototypes are withheld: novel-without-support mIoU rises from 7.05 to 37.02 on ScanNet and from 7.08 to 25.54 on ScanNet200, while base and supported-class mIoU are preserved. Without any support, OW-PC still recovers more novel instances than every baseline, reaching 73.32% instance recall on S3DIS versus 21.12% for GFS-VL. Full-support accuracy remains unchanged in Table 1, and Figure 3 shows that the τ_b/δ_h margins stay within 0.05 points. Appendix D.5 details each regime.

5 Conclusion

We propose a reliability–protection–growth framework for generalized few-shot 3D segmentation. DVD supplies sparse, high-precision anchors; CBG fixes only calibrated base regions; and OW-PC recovers components from the remaining structure with a ddCRP prior and a support-availability conditional posterior. The same inference rule yields named novel instances when support is available and category-agnostic unknown instances when it is not, while protected base elements remain immutable. The framework therefore removes the dependence of novel-class recovery on dense pseudo-labels and support prototypes, and its thresholds are fixed once calibrated on held-out base episodes. Future work will extend this formulation to larger open-world 3D foundation models.

AI USE STATEMENT

No generative AI tools were used to help develop theoretical models or conceptual frameworks, formulate mathematical claims or proofs, propose or refine hypotheses, design research methodology or experiments, generate or clean datasets, perform qualitative analysis, or interpret results. The central ideas, the methodological design, the mathematical formulation, the experimental protocol, and the interpretation of the results are entirely the authors' own work; generative AI assistance was confined to manuscript editing and to parts of the code implementation, as detailed below.

We used generative AI tools for the recommended-disclosure tasks of editing the manuscript to improve readability, formatting the LaTeX layout, and assisting with parts of the code implementation, such as boilerplate routines, refactoring, and debugging; all AI-assisted code was reviewed, tested, and validated by the authors. These tools did not generate or alter any scientific content, claims, or results, and did not contribute to the method, the experiments, or the conclusions. We have reviewed all AI-assisted output and verified that it accurately reflects our own work, and we will take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

Our study uses only publicly available indoor 3D point-cloud benchmarks. We fully comply with the usage specifications and terms of use established by the original authors of these datasets, and we have obtained permission to use them for this research. No human subjects participated in this research, no new data were collected, and the experiments involve no personal or sensitive information. As a general-purpose 3D perception method, our framework may be applied to downstream systems such as robotics, augmented reality, or surveillance; we do not identify any novel harmful capability specific to this work, but we encourage privacy-preserving and responsible deployment. The authors declare no competing interests, and this work was conducted in accordance with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure reproducibility. The complete formulation of DVD, CBG, and OW-PC, together with the derivations of the component posteriors, the acceptance rules, and the end-to-end inference pseudocode are provided in Appendices A and B; the full calibration and evaluation protocols are given in Appendix C. The main text specifies the problem setting, the baselines, the evaluation metrics, and the datasets, and all experiments use the publicly available S3DIS and ScanNet200 benchmarks, whose preprocessing, splits, and evaluation settings we describe. We will publicly release the source code, configuration files, and scripts that reproduce all reported tables and figures.

REFERENCES

Samir Abou Haidar, Alexandre Chariot, Mehdi Darouich, Cyril Joly, and Jean-Emmanuel Deschaud. Are We Ready for Real-Time LiDAR Semantic Segmentation in Autonomous Driving? Workshop on Planning, Perception, and Navigation for Intelligent Vehicles, IROS, Oct 2024. URL https://cea.hal.science/cea-04799429. Poster.

Zhaochong An, Guolei Sun, Yun Liu, Fayao Liu, Zongwei Wu, Dan Wang, Luc Van Gool, and Serge Belongie. Rethinking Few-shot 3D Point Cloud Semantic Segmentation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3996–4006, Los Alamitos, CA, USA, Jun 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00383. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00383.

Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. Generalized few-shot 3d point cloud segmentation with vision-language model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16997–17007, 2025a.

Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation, 2025b. URL https://arxiv.org/abs/2410.22489.

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152 – 1174, 1974. doi: 10.1214/aos/1176342871. URL https://doi.org/10.1214/aos/1176342871.

Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1534–1543, 2016. doi: 10.1109/CVPR.2016.170.

David M. Blei and Peter I. Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(74):2461–2488, 2011. URL http://jmlr.org/papers/v12/blei11a.html.

Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot semantic segmentation. In *Proceedings of the 32nd British Machine Vision Conference*, November 2021.

Christopher Bongsoo Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3070–3079, 2019. URL https://api.semanticscholar.org/CorpusID:121123422.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017. URL https://api.semanticscholar.org/CorpusID:7684883.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973. URL https://api.semanticscholar.org/CorpusID:9777424.

Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks, 2018. URL https://arxiv.org/abs/1711.04043.

Soumya Ghosh, Andrei B. Ungureanu, Erik B. Sudderth, and David M. Blei. Spatial distance dependent chinese restaurant processes for image segmentation. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 1476–1484, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Dan Halperin and Niklas Eisl. Point cloud based scene segmentation: A survey, 2025. URL https://arxiv.org/abs/2503.12595.

Shuting He, Xudong Jiang, Wei Jiang, and Henghui Ding. Prototype adaption and projection for few- and zero-shot 3d point cloud semantic segmentation. *Trans. Img. Proc.*, 32:3199–3211, January 2023. ISSN 1057-7149. doi: 10.1109/TIP.2023.3279660. URL https://doi.org/10.1109/TIP.2023.3279660.

Krishna Murthy Jatavallabhula, Ali Kuwajerwala, Qiao Gu, Mohd. Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Varma Keetha, Ayush Kumar Tewari, Joshua B. Tenenbaum, Celso M. de Melo, M. Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *ArXiv*, abs/2302.07241, 2023. URL https://api.semanticscholar.org/CorpusID:256846496.

Junha Lee, Chunghyun Park, Jaesung Choe, Yu-Chiang Frank Wang, Jan Kautz, Minsu Cho, and Chris Choy. Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14089–14101, 2025. doi: 10.1109/CVPR52734.2025.01315.

Seungjun Lee, Zihan Wang, Yunsong Wang, and Gim Hee Lee. Embodiedsplat: Online feed-forward semantic 3dgs for open-vocabulary 3d scene understanding, 2026. URL https://arxiv.org/abs/2603.04254.

Haosheng Li, Weixin Mao, Weipeng Deng, Chenyu Meng, Rui Zhang, Fan Jia, Tiancai Wang, Haoqiang Fan, Hongan Wang, and Xiaoming Deng. Seggrasp: Zero-shot task-oriented grasping via semantic and geometric guided segmentation, 2024. URL https://arxiv.org/abs/2410.08901.

Qijing Li, Jingxiang Sun, Liang An, Zhaoqi Su, Hongwen Zhang, and Yebin Liu. Semanticsplat: Feed-forward 3d scene understanding with language-aware gaussian fields, 2025. URL https://arxiv.org/abs/2506.09565.

Jie Liu, Jiayi Shen, Pan Zhou, Jan-Jakob Sonke, and Efstratios Gavves. Probabilistic prototype calibration of vision-language models for generalized few-shot semantic segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21155–21165, 2025. doi: 10.1109/ICCV51701.2025.01965.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sungju Hwang, and Yi Yang. LEARNING TO PROPAGATE LABELS: TRANSDUCTIVE PROPAGATION NETWORK FOR FEW-SHOT LEARNING. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyVuRiC5K7.

Phuc 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4018–4028, 2024. doi: 10.1109/CVPR52733.2024.00385.

Frederik Pahde, Mihai Puscas, Jannik Wolff, Tassilo Klein, Nicu Sebe, and Moin Nabi. Low-shot learning from imaginary 3d model. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 978–985, 2019. doi: 10.1109/WACV.2019.00109.

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.

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, 2021. URL https://arxiv.org/abs/2103.00020.

David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 125–141, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19827-4.

Yixiao Song, Qingyong Li, Wen Wang, and Zhicheng Yan. Pointgs: Semantic-consistent unsupervised 3d point cloud segmentation with 3d gaussian splatting, 2026. URL https://arxiv.org/abs/2605.11520.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: open-vocabulary 3d instance segmentation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Xujing Tao, Chuxin Wang, Yubo Ai, Zhixin Cheng, Zhuoyuan Li, Liangsheng Liu, Yujia Chen, Xinjun Li, Qiao Li, Wenfei Yang, and Tianzhu Zhang. Geoguide: Hierarchical geometric guidance for open-vocabulary 3d semantic segmentation, 2026. URL https://arxiv.org/abs/2603.26260.

Vishal Thengane, Zhaochong An, Tianjin Huang, Son Lam Phung, Abdesselam Bouzerdoum, Lu Yin, Na Zhao, and Xiatian Zhu. Scope: Scene-contextualized incremental few-shot 3d segmentation, 2026. URL https://arxiv.org/abs/2603.06572.

Changshuo Wang, Zhijian Hu, Xiang Fang, Zai Yang Yu, Yibin Wu, Mingkun Xu, Yusong Wang, Xingyu Gao, and Prayag Tiwari. Biologically-inspired evolutionary domain symbiosis for few-shot and zero-shot point cloud semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9666–9674, 2026a.

Jiahui Wang, Haiyue Zhu, Haoren Guo, Abdullah Al Mamun, Cheng Xiang, and Tong Heng Lee. Epsegfz: efficient point cloud semantic segmentation for few- and zero-shot scenarios with language guidance. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026b. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i12.37953. URL https://doi.org/10.1609/aaai.v40i12.37953.

Lik Hang Kenny Wong, Xueyang Kang, Kaixin Bai, and Jianwei Zhang. A survey of robotic navigation and manipulation with physics simulators in the era of embodied ai. *ArXiv*, abs/2505.01458, 2025. URL https://api.semanticscholar.org/CorpusID:278327440.

Yating Xu, Conghui Hu, Na Zhao, and Gim Hee Lee. Generalized few-shot point cloud segmentation via geometric words. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21449–21458, 2023. doi: 10.1109/ICCV51070.2023.01966.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *ArXiv*, abs/2101.06395, 2021. URL https://api.semanticscholar.org/CorpusID:231632202.

Canyu Zhang, Zhenyao Wu, Xinyi Wu, Ziyu Zhao, and Song Wang. Few-shot 3d point cloud semantic segmentation via stratified class-specific attention based transformer network. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i3.25449. URL https://doi.org/10.1609/aaai.v37i3.25449.

Zihao Zhang, Aming Wu, Yang Li, Yahong Han, and Jialie Shen. Geometric-aware hypergraph reasoning for novel class discovery in point cloud segmentation, 2026. URL https://arxiv.org/abs/2606.07280.

Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8869–8878, 2021. doi: 10.1109/CVPR46437.2021.00876.

APPENDIX CONTENTS

A Additional Methodological Derivations **15**

A.1 DVD Proposal and Reliability 15

A.2 CBG and Component Evidence 15

A.3 Support Evidence and OW-PC Decoding 15

A.4 Support-Free OW-PC 16

A.5 Novel-Class Cardinality Decoding 18

A.6 Boundary Handling in CBG and OW-PC 18

B End-to-End Framework Procedure **20**

B.1 Inputs, Calibration, and Outputs 20

B.2 Step-by-Step Query Transformation 21

B.3 Inference Algorithm 22

C Detailed Experimental Settings **22**

C.1 Datasets and Class Partitions 22

C.2 Few-Shot Episodes and Registration Protocol 23

C.3 Partial and Support-Free Conditions 23

C.4 Model, Optimization, and Data Processing 24

C.5 Calibration and Leakage Controls 24

C.6 Calibration Protocol Details 24

C.7 Metrics and Aggregation 25

D Further Experimental Analysis **25**

D.1 Multi-Seed Stability 25

D.2 Recovery Beyond Pointwise VLM Predictions 25

D.3 Detailed ScanNet200 Recall 26

D.4 ScanNet200 Partial-Prototype Class Breakdown 26

D.5 Support-Availability Details 27

D.6 Primary–Auxiliary VLM Pairings 28

D.7 Sensitivity, Support Quality, and Efficiency 29

E More Qualitative Visualizations **29**

APPENDIX

A ADDITIONAL METHODOLOGICAL DERIVATIONS

This appendix gives the definitions omitted from the main paper. All calibration parameters are learned on held-out base episodes and then fixed for novel evaluation.

A.1 DVD PROPOSAL AND RELIABILITY

For support instance k, let $\bar{f}_c^{(k)}$ be the mean labeled feature of class c. DVD uses the episode-specific compatibility and initial proposal odds

$$s_i^c = \exp\left(-\frac{\min_k \|f_i - \bar{f}_c^{(k)}\|_2^2}{2\sigma_s^2}\right), \qquad L_i^{c,0} = b_0 + \lambda_{\text{pri}} \operatorname{logit}_\delta(p_i^{\text{pri}}(c)) + \lambda_S \operatorname{logit}_\delta(s_i^c), \quad (22)$$

where $\operatorname{logit}_\delta(q) = \log[(q+\delta)/(1-q+\delta)]$. Let u_i^c describe whether the auxiliary prediction matches c, another novel class, a base class, or neither. On pseudo-novel base episodes, we estimate $\psi_u^+ = P(u \mid z_i^c = 1)$ and $\psi_u^- = P(u \mid z_i^c = 0)$, giving

$$\Lambda_i^{\text{aux},c} = \frac{\psi_{u_i^c}^+ + \delta}{\psi_{u_i^c}^- + \delta}, \qquad q_i^*(z_i^c = 1) = \operatorname{sigmoid}\Big(L_i^{c,0} + \lambda_{\text{aux}} \log \Lambda_i^{\text{aux},c}\Big). \quad (23)$$

This posterior is the stationary point of the variational objective

$$\min_{q_i} \mathcal{J}(q_i) = D_{\text{KL}}\big(q_i(z_i^c) \,\big\|\, q_i^0(z_i^c)\big) - \lambda_{\text{aux}} \mathbb{E}_{q_i}\big[\log P(u_i^c \mid z_i^c)\big], \quad (24)$$

which balances fidelity to the episodic proposal against auxiliary verification energy. The retained anchors act as clamped semantic observations in OW-PC, while CBG protects only calibrated base elements and the exposed domain remains available for component-level recovery. Anchor precision, base stability, and scene-level recall are therefore controlled by separate variables rather than traded through one dense pseudo-label threshold. DVD retains i as class-c anchor only if $c = \arg\max_{c'} R_i^{c'}$ and $q_i^*(z_i^c = 1) > \tau_R$.

A.2 CBG AND COMPONENT EVIDENCE

The calibrated two-view score, base margin, and protected set are

$$\begin{aligned} G_i(\ell) &= \log(p_i^{\text{pri}}(\ell) + \delta) + \lambda_B \log(p_i^{\text{aux}}(\ell) + \delta), \\ g_i &= G_i(\hat{b}_i) - \max_{\ell \neq \hat{b}_i} G_i(\ell), \\ \mathcal{P}_B &= \{i : \hat{y}_i^{\text{pri}} = \hat{y}_i^{\text{aux}} = \hat{b}_i,\ g_i > \tau_b\}. \end{aligned} \quad (25)$$

The exposed domain is $\mathcal{E} = Q \setminus \mathcal{P}_B$. For component $A \subseteq \mathcal{E}$, the strongest base counterfactual is

$$\hat{b}_A = \arg\max_{b \in \mathcal{C}_{\text{base}}} \frac{1}{|A|} \sum_{i \in A} G_i(b), \qquad e_A^B = \frac{1}{|A|} \sum_{i \in A} G_i(\hat{b}_A). \quad (26)$$

The label $\hat{b}_A$ is output whenever the component is assigned the base state. Hence a base-to-novel revision is always evaluated against the strongest coherent base explanation, rather than a pointwise score.

A.3 SUPPORT EVIDENCE AND OW-PC DECODING

For each available class $c \in \mathcal{C}_S$, normalized support features $F_c = \{f_{cn}\}_{n=1}^{N_c}$ are clustered into at most $M_{\max}$ fixed modes by farthest-point initialization followed by spherical assignment to normalized centroids μ_{cm}, with mode weights $\pi_{cm} = n_{cm}/N_c$ and dispersion-based temperatures T_{cm} (Equation 11). The named component evidence is

$$e_A^N(c) = \log \sum_{m=1}^{M_c} \pi_{cm} \exp\left(\frac{1}{|A| T_{cm}} \sum_{i \in A} f_i^\top \mu_{cm}\right). \quad (27)$$

The local links induce $\Pi(a)$. The available hypotheses and their energies are

$$\mathcal{H}(S) = \mathcal{C}_S \cup \{\mathrm{B}, \mathrm{bg}\} \cup \{\mathrm{U} : \xi_U = 1\},$$
$$E_A(h; S) = \begin{cases} e_A^N(c), & h = c \in \mathcal{C}_S, \\ e_A^U, & h = \mathrm{U}, \\ e_A^B, & h = \mathrm{B}, \\ e_A^{\mathrm{bg}}, & h = \mathrm{bg}, \end{cases} \tag{28}$$

where $e_A^U = \lambda_{\mathrm{coh}} \operatorname{coh}(A) - \lambda_\partial \operatorname{cost}_\partial(A)$. If C_A is the set of anchor labels in A, its feasible set is

$$\mathcal{H}_A(S) = \begin{cases} \{c\}, & C_A = \{c\}, \\ \emptyset, & |C_A| > 1, \\ \mathcal{H}(S), & C_A = \emptyset. \end{cases}$$

Thus conflicting anchors have zero posterior mass. We define the calibrated background evidence as

$$e_A^{\mathrm{bg}} = \frac{1}{|A|} \sum_{i \in A} G_i(\mathrm{bg}),$$

where $G_i(\mathrm{bg})$ is the frozen model's calibrated background score (or calibrated complement score when no explicit background logit is supplied).

After partition inference, OW-PC decodes $\hat{h}_A = \arg\max_{h \in \mathcal{H}_A(S)} \{\log \pi_h + E_A(h; S)\}$. A named or unknown component is emitted only when

$$\begin{aligned} |A| &\geq n_{\min}(\hat{h}_A), \\ \mathrm{LBF}_A(\hat{h}_A) &= \log \pi_{\hat{h}_A} + E_A(\hat{h}_A; S) \\ &\quad - \log \sum_{h \in \mathcal{H}_A(S) \setminus \{\hat{h}_A\}} \pi_h \exp(E_A(h; S)) \geq \delta_{\hat{h}_A}. \end{aligned} \tag{29}$$

Protected elements are never modified. If an exposed component fails the size or LBF test, it falls back to the higher prior-weighted base/background state. A base fallback outputs $\hat{b}_A$; a background fallback outputs bg.

A.4 Support-Free OW-PC

OW-PC is defined by one support-availability conditional posterior. The support-free case is obtained by setting $\mathcal{C}_S = \emptyset$ and $\xi_U = 1$, rather than replacing OW-PC with a separate clustering method. DVD produces no class-bearing anchors and no support-mode evidence is available, but CBG still defines the exposed domain $\mathcal{E}$ and supplies the base counterfactual. The same local link prior proposes candidate partitions, while the OW-PC component marginal decides whether each proposed structure is better explained as unknown, base, or background. The explanation priors π_h, minimum sizes $n_{\min}(h)$, and margins δ_h are calibrated on held-out base episodes and then fixed; no query label enters calibration.

A.4.1 Unknown-Component State and Evidence

Without support, the available explanation set is

$$\mathcal{H}_{\mathrm{sf}} = \{\mathrm{U}, \mathrm{B}, \mathrm{bg}\}, \tag{30}$$

where U denotes a category-agnostic unknown object, B the strongest calibrated base explanation, and bg background. The unknown-component evidence is

$$e_A^U = \lambda_{\mathrm{coh}} \operatorname{coh}(A) - \lambda_\partial \operatorname{cost}_\partial(A). \tag{31}$$

Here $\operatorname{coh}(A)$ aggregates the spatial–feature compatibility of links internal to A, while $\operatorname{cost}_\partial(A)$ penalizes links that cross strong geometric or feature discontinuities. Both terms are computed solely from the query coordinates and frozen features. Neither a support feature nor a semantic class label enters this score.

The support-free component energy is

$$E_A^{\mathrm{sf}}(h) = \begin{cases} e_A^U, & h = \mathrm{U}, \\ e_A^B, & h = \mathrm{B}, \\ e_A^{\mathrm{bg}}, & h = \mathrm{bg}. \end{cases} \tag{32}$$

Unlike the support-conditioned energy, Equation 32 contains no named state c, support-mode evidence $e_A^N(c)$, or anchor-feasibility restriction.

A.4.2 Marginal Posterior and Coordinate Update

Let $\Pi(a)$ be the components induced by the local links and let π_h be the calibrated prior for explanation h. OW-PC first marginalizes the support-free explanations within each component,

$$\Psi_{\mathrm{sf}}(A) = \sum_{h \in \mathcal{H}_{\mathrm{sf}}} \pi_h \exp\big(E_A^{\mathrm{sf}}(h)\big). \tag{33}$$

Combining these component marginals with the ddCRP link prior gives the support-free OW-PC posterior

$$p_{\mathrm{sf}}(a \mid Q) \propto \prod_{i \in \mathcal{E}} \phi_i(a_i) \prod_{A \in \Pi(a)} \Psi_{\mathrm{sf}}(A). \tag{34}$$

Taking logarithms yields the coordinate-MAP objective

$$\mathcal{J}_{\mathrm{sf}}(a) = \sum_{i \in \mathcal{E}} \log \phi_i(a_i) + \sum_{A \in \Pi(a)} \log \Psi_{\mathrm{sf}}(A) + \mathrm{const}. \tag{35}$$

Suppose the link of element i changes from j to j', replacing the affected component set $\mathcal{A}_{\mathrm{old}}$ by $\mathcal{A}_{\mathrm{new}}$. All unaffected factors cancel, so the update score is

$$\Delta\mathcal{J}_i(j \to j') = \log \frac{\phi_i(a_i = j')}{\phi_i(a_i = j)} + \sum_{A \in \mathcal{A}_{\mathrm{new}}} \log \Psi_{\mathrm{sf}}(A) - \sum_{A \in \mathcal{A}_{\mathrm{old}}} \log \Psi_{\mathrm{sf}}(A). \tag{36}$$

The update is accepted when $\Delta\mathcal{J}_i > 0$. Because a link change can merge or split components, the last two terms recompute unknown-object coherence, boundary cost, and base/background counterfactuals. Crucially, the update uses the log-sum-exp marginal rather than a preselected component label. After the partition converges, the component posterior and decoded state are

$$p_{\mathrm{sf}}(h \mid A, Q) = \frac{\pi_h \exp(E_A^{\mathrm{sf}}(h))}{\Psi_{\mathrm{sf}}(A)}, \qquad \hat{h}_A = \arg\max_{h \in \mathcal{H}_{\mathrm{sf}}} \{\log \pi_h + E_A^{\mathrm{sf}}(h)\}. \tag{37}$$

A.4.3 Unknown-Component Acceptance Rule

For a candidate component A, the unknown explanation is compared with the marginalized base and background counterfactuals:

$$\mathrm{LBF}_A(\mathrm{U}) = \log \pi_{\mathrm{U}} + e_A^U - \log\Big(\pi_{\mathrm{B}} \exp(e_A^B) + \pi_{\mathrm{bg}} \exp(e_A^{\mathrm{bg}})\Big). \tag{38}$$

OW-PC reports A as unknown only if $\hat{h}_A = \mathrm{U}$ and

$$|A| \geq n_{\min}^U, \qquad \mathrm{LBF}_A(\mathrm{U}) \geq \delta_U. \tag{39}$$

Otherwise, the component is not emitted as unknown and retains the applicable base or background outcome. The thresholds $n_{\min}^U$ and δ_U, the explanation priors, and the evidence weights in Equation 31 are calibrated on held-out base episodes and fixed before evaluation.

A.4.4 Output Semantics and Evaluation

Support-free OW-PC outputs base, background, and unknown components without category names. Ground-truth class labels are consulted only after inference to evaluate the recovered components. Instance recall at IoU 0.50 counts the fraction of ground-truth instances whose best-overlapping predicted unknown component reaches IoU 0.50; mean best IoU averages that best overlap over

all ground-truth instances. This post-hoc matching changes neither the inferred partition nor the decoded component states.

To isolate the role of semantic pseudo-labels, we reuse the legacy OpenScene SparseUNet feature extractor and checkpoint used by GFS-VL, while providing OW-PC with no class names, semantic logits, or support prototypes. The support-free instance-recovery results in the main paper show that OW-PC can use the frozen feature representation for structural discovery without consuming novel-class pseudo-labels or classifier outputs.

Figure 4b gives qualitative results under the support-free setting. The recovered components are evaluated as category-agnostic unknown objects because no labeled support prototype or novel-class text is supplied.

Without semantic support, GFS-VL leaves substantial portions of the bathtub, plant, kitchen counter, nightstand, and shower curtain unrecovered. OW-PC recovers a larger fraction of each target extent and produces predictions that more closely follow the corresponding ground-truth component. This improvement is particularly visible for the plant and nightstand examples, where the baseline covers only a fragment of the object.

A.5 Novel-Class Cardinality Decoding

Besides point and component labels, the accepted components provide a dataset-level estimate of how many novel semantic categories are recovered. The protocol has two stages. OW-PC first applies its ddCRP prior to the exposed domain $\mathcal{E}$ (Equation 13) and retains the components whose log-Bayes factor passes the acceptance test (Equation 20); these components define an initial set of category-agnostic structural classes. Because every exposed element may link to itself, new components are created as needed and retained only when the acceptance test passes, so the number of components is inferred from the scene rather than fixed in advance. Each accepted component is then mapped to a semantic category through the support-availability conditional explanation $\hat{h}_A = \arg\max_{h \in \mathcal{H}_A(S)} \{\log \pi_h + E_A(h; S)\}$ (Equation 16), and components decoded as base, background, or unknown contribute no novel category. Let $\mathcal{A}_{\mathrm{N}}$ be the accepted components decoded as named novel classes. We report the deduplicated set and its cardinality as

$$\widehat{\mathcal{C}}_{\mathrm{novel}}(Q) = \{\hat{h}_A : A \in \mathcal{A}_{\mathrm{N}},\ \hat{h}_A \in \mathcal{C}_S\}, \qquad \widehat{K}_{\mathrm{novel}}(Q) = \left|\widehat{\mathcal{C}}_{\mathrm{novel}}(Q)\right|. \tag{40}$$

Each experiment is repeated five times, and we report the arithmetic mean of the decoded cardinality across these runs.

Multiple spatial components assigned to the same named class therefore contribute once, whereas a support class with no accepted component is not counted. The table reports the ground-truth category count and the five-run average of the decoded cardinality. This category-level measure is distinct from instance recall and does not penalize fragmentation of one category into disconnected instances. In support-free episodes, OW-PC instead returns the number of accepted unknown components, preserving the distinction between category cardinality and instance discovery.

A.6 Boundary Handling in CBG and OW-PC

A.6.1 Boundary Treatment in the Current Model

The current framework handles boundaries implicitly through three mechanisms. First, CBG removes protected base elements from the exposed domain, so OW-PC cannot link through or relabel a calibrated base point. This creates a hard boundary around reliable base regions. Second, the local ddCRP prior decays with both spatial and feature distance, making a link across a geometric gap or feature discontinuity less probable. Third, every proposed merge changes $\Psi(A; S)$ by marginalizing all feasible named, unknown, base, and background explanations. A link that absorbs floor or table points into a stool component can therefore weaken the marginal component evidence and the decoded explanation's LBF, causing OW-PC to reject the merge or the completed component.

These mechanisms provide boundary awareness but not explicit boundary supervision. If two touching objects have similar frozen features, the current kernel may still assign excessive probability to a cross-object link. We therefore formulate an explicit boundary-aware extension below. This exten-

sion is presented separately from the main model and will be validated by the experiments described in Section A.6.5.

A.6.2 Three-Way Base–Novel–Background Boundaries

The framework does not predict the three boundary types with independent binary heads. Instead, local links determine where a component may end, while CBG and the component explanations determine which semantic state lies on either side of that boundary. This yields the following treatment of the three pairwise interfaces.

Base–novel boundary. For a calibrated base element in $\mathcal{P}_B$, the boundary is hard: the element is excluded from OW-PC and cannot be absorbed by a novel or unknown component. For an ambiguous base prediction in $\mathcal{E}$, the boundary remains revisable. A neighboring component may include that element only when the resulting partition improves the marginal OW-PC posterior; a decoded named or unknown explanation must then pass its LBF threshold against the base and all other available counterfactuals. Thus CBG protects the reliable interior of a base object, while posterior component competition resolves its uncertain fringe.

Base–background boundary. Protected base elements again remain fixed. For an exposed component, OW-PC includes both e_A^B and the calibrated background score e_A^{bg} in $\Psi(A; S)$. After partition inference, the component is decoded as base or background according to its prior-weighted energy when neither a named nor an unknown explanation is accepted. This prevents low-confidence clutter adjacent to a base object from being preserved merely because individual points received weak base probabilities.

Novel–background boundary. The outer extent of a named or unknown component is determined jointly by the local partition prior and the marginal competition in $\Psi(A; S)$. Adding compatible object points can strengthen support-conditioned evidence or unknown-object coherence, whereas absorbing floor, wall, or isolated clutter generally weakens those explanations and may introduce high-cost links. After partition inference, a selected named or unknown state h is accepted only when $|A| \geq n_{\min}(h)$ and $\mathrm{LBF}_A(h) \geq \delta_h$. The same rule applies to anchored and unanchored components; anchors restrict feasible named identities but do not bypass component acceptance.

Support-conditional component decision. Let C_A be the set of anchor classes in component A. After marginal partition inference proposes the component boundary, OW-PC decodes

$$\hat{h}_A = \arg \max_{h \in \mathcal{H}_A(S)} \{\log \pi_h + E_A(h; S)\}. \tag{41}$$

For $\hat{h}_A \in \mathcal{C}_S \cup \{\mathrm{U}\}$, the decoded state becomes a completed semantic target exactly when

$$|A| \geq n_{\min}(\hat{h}_A), \qquad \mathrm{LBF}_A(\hat{h}_A) \geq \delta_{\hat{h}_A}. \tag{42}$$

All experiments are conducted on two NVIDIA RTX 4090 GPUs using mixed-precision (FP16) training. Components containing anchors from multiple named classes are infeasible. If an anchored component fails Equation 42, its anchor observations remain sparse supervision but its proposed non-anchor extent is not converted into a dense target. Equations 41–42 make the division of responsibility explicit: the local prior proposes boundaries, CBG fixes reliable base interiors, and OW-PC marginalizes the available component explanations before decoding and acceptance.

A.6.3 Boundary-Aware Link Potential

Let n_i be a local surface normal estimated from query coordinates, and let $\widetilde{G}_i$ be the normalized calibrated semantic-evidence vector. For neighboring exposed elements (i, j), define the boundary strength

$$b_{ij} = \eta_n\big(1 - |n_i^\top n_j|\big) + \eta_f \frac{\|f_i - f_j\|_2^2}{2\sigma_f^2} + \eta_G \|\widetilde{G}_i - \widetilde{G}_j\|_1, \tag{43}$$

where the three terms capture surface-orientation change, feature discontinuity, and calibrated semantic disagreement. The boundary-aware link potential is

$$\phi_i^{\partial}(a_i = j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma_x^2} - \frac{\|f_i - f_j\|_2^2}{2\sigma_f^2} - \lambda_{\partial} b_{ij} \right), \tag{44}$$

for $j \in \mathcal{N}(i) \cap \mathcal{E}$, while the self-link remains α. Links to $j \in \mathcal{P}_B$ are infeasible. Consequently, a strong geometric or semantic discontinuity suppresses cross-boundary growth even when two points are spatially close.

A.6.4 Partition-Level Boundary Consistency

Pairwise suppression alone can over-segment noisy surfaces. To balance boundary preservation and within-object connectivity, let $\mathcal{E}_{\mathrm{in}}(A)$ denote neighbor edges internal to component A and $\mathcal{E}_{\mathrm{cut}}(A)$ denote edges crossing from A to another exposed component. After mapping b_{ij} to $\bar{b}_{ij} \in [0, 1]$, define

$$C_{\partial}(A) = \frac{1}{|\mathcal{E}_{\mathrm{in}}(A)|} \sum_{(i,j) \in \mathcal{E}_{\mathrm{in}}(A)} (1 - \bar{b}_{ij}) + \frac{1}{|\mathcal{E}_{\mathrm{cut}}(A)|} \sum_{(i,j) \in \mathcal{E}_{\mathrm{cut}}(A)} \bar{b}_{ij}, \tag{45}$$

with an empty average defined as zero. The extended component energy is

$$E_A^{\partial}(h; S) = E_A(h; S) + \lambda_C C_{\partial}(A). \tag{46}$$

The first term in $C_{\partial}(A)$ rewards smooth internal connectivity, while the second rewards cuts aligned with strong boundaries. The extended marginal is obtained by replacing $E_A(h; S)$ with $E_A^{\partial}(h; S)$ inside $\Psi(A; S)$. Because $C_{\partial}(A)$ is independent of h, it changes the partition score but cancels when comparing explanations within a fixed component. The minimum-size and LBF rules remain unchanged and prevent the boundary reward from accepting tiny unsupported components.

A.6.5 Boundary Evaluation Protocol

We will compare the current spatial-feature kernel, the normal-only extension, the semantic-discontinuity extension, and the complete boundary-aware posterior. Evaluation will include boundary F-score at multiple voxel tolerances, trimap IoU near ground-truth boundaries, base-to-novel leakage, novel-to-base omission, component fragmentation, and the standard base/novel mIoU and HM. Results will be stratified into touching-object and separated-object subsets. The touching subset is essential because global mIoU can hide leakage precisely where a novel stool contacts a base table or the floor.

We will additionally visualize accepted and rejected links around contact regions. This will show whether an improvement is caused by better boundary localization or simply by more conservative component acceptance. All boundary hyperparameters will be calibrated on held-out base episodes under the same protocol used for DVD and CBG.

B End-to-End Framework Procedure

This section describes how a query scene is transformed from raw model outputs into accepted component labels. DVD identifies trustworthy novel-class observations, CBG determines which base predictions must remain fixed, and OW-PC completes the remaining query structure by jointly evaluating candidate partitions and their available semantic explanations. The ddCRP is used only as OW-PC's local partition prior; semantic completion is governed by the support-availability conditional component posterior.

B.1 Inputs, Calibration, and Outputs

All experiments are conducted on two NVIDIA RTX 4090 GPUs using mixed-precision (FP16) training. For each episode, the input consists of an unlabeled query scene $Q = \{(x_i, f_i, p_i^{\mathrm{pri}}, p_i^{\mathrm{aux}})\}_{i=1}^N$ and a possibly incomplete collection of labeled supports $\{S_c^{(k)}\}_{c \in \mathcal{C}_S}$. Here x_i is the voxel coordinate, f_i is the frozen language-aligned query feature, p_i^{pri} and p_i^{aux} are the primary and auxiliary VLM class distributions, and $\mathcal{C}_S$ contains exactly the novel classes with available

support. A switch ξ_U enables the category-agnostic unknown hypothesis when support is absent or partial. Held-out base-class episodes are converted into pseudo-novel episodes to estimate DVD's smoothed relation likelihoods and calibrate the anchor threshold, base-protection threshold, explanation priors, component-size thresholds, and log-Bayes-factor thresholds. Novel query labels are never used for calibration.

OW-PC uses the same inference rule in every support regime. Full support yields named-novel, base, and background outcomes; partial support additionally permits unknown components for unsupported classes; and zero support yields category-agnostic unknown, base, and background outcomes. Protected base elements preserve their original labels. Elements in an accepted named or unknown component receive the decoded component state, while rejected exposed elements retain the applicable base or background outcome under the evaluation protocol.

B.2 Step-by-Step Query Transformation

Step 1: construct episode-specific novel proposals. For each query element i and novel class c, DVD combines the primary novel probability $p_i^{\mathrm{pri}}(c)$ with the similarity s_i^c to the labeled support instances. This produces an initial log-odds $L_i^{c,0}$. At this point the candidates can be dense and noisy: some cover only fragments of a true instance, while others occur on base objects or background.

Step 2: distill sparse high-precision anchors. The auxiliary VLM does not directly provide pseudo-ground-truth labels. Instead, its prediction is converted into a calibrated semantic relation to candidate class c: match, another novel class, a base class, or other. The corresponding likelihood ratio updates $L_i^{c,0}$ to the posterior reliability R_i^c. Thresholding this posterior converts the noisy proposal map into a sparse anchor set $\mathcal{S}$. Each retained anchor carries a novel class identity and a reliability weight, but the unselected query points remain available for later structural inference.

Step 3: protect only reliable base elements. CBG combines the two semantic views into calibrated evidence $G_i(\ell)$. A query element enters the protected set $\mathcal{P}_B$ only if the two views agree on the same base class and that class has sufficient margin over every base and novel alternative. These protected elements are removed from structural growth and form a hard base-preservation constraint. All other elements form the exposed domain $\mathcal{E} = Q \setminus \mathcal{P}_B$. Exposure does not itself change a label; it only permits reconsideration by component evidence.

Step 4: instantiate the available component hypotheses. For every $c \in \mathcal{C}_S$, normalized labeled support features are clustered into at most $M_{\max}$ observed modes. Each mode stores its centroid μ_{cm}, empirical mass π_{cm}, and uncertainty temperature T_{cm}; sparse or dispersed modes therefore provide less decisive evidence. These modes define $e_A^N(c)$ for named hypotheses and are never updated from query components. OW-PC then forms $\mathcal{H}(S) = \mathcal{C}_S \cup \{\mathrm{B}, \mathrm{bg}\}$ and adds U when $\xi_U = 1$, so the hypothesis set reflects support availability without changing structural inference.

Step 5: infer the partition by marginal component evidence. Every exposed element links either to a spatial neighbor or to itself. Spatially close and feature-compatible pairs receive larger dd-CRP link potential, while the self-link mass allows the number of components to adapt to the scene. For each induced component A, anchors restrict the feasible set $\mathcal{H}_A(S)$, and OW-PC marginalizes all feasible named, unknown, base, and background explanations through $\Psi(A; S)$. A coordinate link update is retained only when the change in its local link prior plus the affected $\log \Psi(A; S)$ terms improves the posterior. Component labels are therefore not fixed during partition inference, and components containing conflicting anchor classes have zero posterior mass.

Step 6: decode and accept completed components. After partition inference, OW-PC decodes $\hat{h}_A$ from the largest prior-weighted component energy. A named or unknown component is accepted only if it satisfies its state-specific minimum size $n_{\min}(\hat{h}_A)$ and its log-Bayes factor against all alternative explanations exceeds $\Delta_{\hat{h}_A}$. Anchors preserve a named class through the feasible-set constraint, but the non-anchor extent of an anchored component is still subject to this acceptance rule. An unanchored component can be decoded as a supported named class or, when enabled, as U. Thus recovery depends on component evidence against all counterfactuals rather than propagation from a pointwise pseudo-label.

Algorithm 1 DVD–CBG–OW-PC inference for one generalized few-shot episode.

Require: Available supports $\{S_c^{(k)}\}_{c\in\mathcal{C}_S}$, query elements Q, frozen features and distributions $(f_i, p_i^{\mathrm{pri}}, p_i^{\mathrm{aux}})$, unknown switch ξ_U, calibrated priors and thresholds
Ensure: Protected base labels and accepted named or unknown components

```
Construct support-conditioned proposal log-odds L_i^{c,0} for every c ∈ C_S
Convert the auxiliary prediction into relation state u_i^c and likelihood ratio Λ_i^{aux,c}
Compute R_i^c = L_i^{c,0} + λ_aux log Λ_i^{aux,c} and retain sparse anchors S
Compute calibrated class evidence G_i(ℓ) and base margin g_i
Form protected base set P_B and exposed domain E = Q \ P_B
for each supported class c ∈ C_S do
    Cluster labeled support features into fixed modes (μ_cm, π_cm, T_cm)
end for
Form H(S) from available named, base, background, and unknown hypotheses
Initialize every exposed element with a self-link
repeat
    for each exposed element i ∈ E do
        Form the affected components for each feasible self-link or neighbor link
        Evaluate the link-prior change and the affected log Ψ(A; S) terms
        Select the feasible link with the largest posterior improvement
    end for
until no link update improves the marginal OW-PC posterior
for each induced component A ∈ Π(a) do
    Decode ĥ_A = arg max_{h∈H_A(S)} {log π_h + E_A(h; S)}
    Accept named or unknown states using n_min(ĥ_A) and LBF_A(ĥ_A) ≥ δ_{ĥ_A}
    For rejected components, output b̂_A for the higher prior-weighted base fallback, otherwise
    output background
end for
Combine accepted components with the immutable labels in P_B
```

B.3 Inference Algorithm

Training targets. DVD anchors supervise the segmentation model with posterior reliability weights. Accepted OW-PC components supply component-confidence-weighted targets. Support modes remain fixed throughout this process, and no unlabeled query feature is written back into the support representation.

C Detailed Experimental Settings

This section gives the complete dataset, class-partition, episode, support-availability, training, and evaluation protocol used by the experiments. We use the generalized few-shot convention throughout: base classes have training annotations, novel classes are excluded from base-class training, and a query scene is evaluated jointly over the base and novel vocabularies. All class lists below are ordered lists; the order is also the classifier and evaluation order. Background or ignored labels are excluded from the class-wise IoU averages.

C.1 Datasets and Class Partitions

We use the official train/validation splits and the standard preprocessing of S3DIS, ScanNet, ScanNet++, and ScanNet200. The base–novel partitions are fixed before any experiment and are never changed between the standard, partial-prototype, and support-free evaluations. Table 6 summarizes the number of classes and the exact base vocabulary; "partial support" counts the novel classes whose support prototypes are retained in the partial-prototype setting, normalized by the total number of novel classes. The novel vocabulary is the complement of the base vocabulary in the ordered

Table 6: Benchmark base–novel partitions.

Dataset	Base	Novel	Total	Partial support	Base classes
ScanNet	13	6	19	3/6	refrigerator, desk, curtain, sofa, bookshelf, bed, table, window, cabinet, door, chair, floor, wall
ScanNet200	12	45	57	22/45	refrigerator, desk, curtain, bookshelf, bed, table, window, cabinet, door, chair, floor, wall
ScanNet++	12	18	30	9/18	wall, floor, door, ceiling, table, window, box, ceiling lamp, light switch, cabinet, chair, heater
S3DIS	10	3	13	1/3	ceiling, floor, wall, beam, column, window, door, table, chair, bookcase

benchmark label list; the full novel lists are given immediately below the table for unambiguous reproduction.

The six ScanNet novel classes are *sink*, *toilet*, *bathtub*, *shower curtain*, *picture*, and *counter*. The 45 ScanNet200 novel classes are *trash can*, *ceiling*, *doorframe*, *object*, *shelf*, *sink*, *picture*, *backpack*, *couch*, *box*, *pillow*, *radiator*, *mirror*, *whiteboard*, *lamp*, *toilet*, *book*, *monitor*, *towel*, *tv*, *clothes*, *coffee table*, *office chair*, *nightstand*, *bag*, *dresser*, *toilet paper*, *recycling bin*, *kitchen cabinet*, *bathtub*, *telephone*, *plant*, *stool*, *keyboard*, *shoe*, *jacket*, *shower curtain*, *armchair*, *microwave*, *computer tower*, *bathroom vanity*, *kitchen counter*, *shower wall*, *paper towel dispenser*, and *file cabinet*. The 18 ScanNet++ novel classes are *monitor*, *whiteboard*, *office chair*, *bottle*, *doorframe*, *keyboard*, *window frame*, *mouse*, *paper*, *blinds*, *trash can*, *telephone*, *book*, *shelf*, *sink*, *windowsill*, *bag*, and *smoke detector*. The three S3DIS novel classes are *sofa*, *board*, and *clutter*.

C.2 Few-Shot Episodes and Registration Protocol

We evaluate both 1-shot and 5-shot episodes. In a K-shot episode, each novel class c provides K labeled support point-cloud instances $\{S_c^{(k)}\}_{k=1}^{K}$, while the query scenes are unlabeled during inference. Support features are normalized and retained as fixed evidence; query features never update a support prototype. The standard setting supplies support for every novel class. The registration protocol repeats the support registration with five fixed registration seeds, denoted `regis1` through `regis5`. A reported five-registration number is the arithmetic mean of the five per-registration class IoU vectors before computing the displayed group means. This is different from averaging already rounded scalar summaries.

If a registration run is interrupted, we report only the completed checkpoint(s) and label the result as a single-checkpoint or partial-registration result. In particular, the ScanNet200 GFS-VL 1-shot partial-prototype checkpoint reported in Table 5 and evaluated in Table 11 uses `regis1`; it is not silently treated as a five-registration average. The final model for each registration is the last checkpoint at the prescribed evaluation epoch (`model_last.pth`).

C.3 Partial and Support-Free Conditions

The partial-prototype condition keeps the base/novel partition unchanged but withholds the support prototype for a fixed subset of novel classes. The retained classes are the first entries of the benchmark's canonical novel-class order, with counts $(3, 22, 9, 1)$ for ScanNet, ScanNet200, ScanNet++, and S3DIS, respectively. For ScanNet200, the supported classes are *trash can*, *ceiling*, *doorframe*, *object*, *shelf*, *sink*, *picture*, *backpack*, *couch*, *box*, *pillow*, *radiator*, *mirror*, *whiteboard*, *lamp*, *toilet*, *book*, *monitor*, *towel*, *tv*, *clothes*, and *coffee table*; the remaining 23 novel classes have no support prototype. Unsupported class names are used only for post-hoc scoring and are not supplied to OW-PC as named hypotheses. The unknown hypothesis is enabled, so the same posterior can choose a named supported class or a category-agnostic unknown component.

The support-free condition sets the available novel support set to the empty set and provides no novel class names, text prompts, semantic logits, or prototypes to the open-world decoder. OW-PC then uses the hypothesis set $\{\mathrm{U}, \mathrm{B}, \mathrm{bg}\}$, where U is category-agnostic unknown. Ground-truth novel names are used only after inference to compute instance metrics; they cannot affect partitioning, decoding, or calibration.

C.4 Model, Optimization, and Data Processing

The segmentation model uses the PT-v3m1 backbone with six input channels (color and normal features) and a 64-dimensional decoder output. The language-aligned 3D VLM feature extractor is the SparseUNet-636 checkpoint. The auxiliary image-language view uses the local ViT-L/14 model at 336-pixel resolution, and the active prompt condition is dataset-specific (ScanNet, ScanNet++, or S3DIS; ScanNet200 follows the ScanNet condition). DVD, CBG, and OW-PC are enabled together for the full model.

The software environment is Python 3.12 with PyTorch 2.8 built against CUDA 12.8; all modules are implemented in this stack. For the reported 1-shot registration runs, training uses two NVIDIA RTX 4090 GPUs, total batch size 2, gradient accumulation of 6, eight data-loader workers per process, seed 42, and mixed-precision FP16. The point budget is 102,400 points per random training crop. The model is optimized with Adam at learning rate 10^{-3} for 20 epochs; evaluation is performed at epoch 20. The training augmentation consists of center shifting, random dropout, rotations, scaling, flipping, coordinate jitter, elastic distortion, color contrast/translation/jitter, voxelization at grid size 0.02, random spherical cropping, color normalization, and tensor conversion. These transformations are applied only to training crops; validation uses deterministic voxelization at the same grid size, full-scene fragments, and inverse mapping to the original points.

Evaluation is single-GPU with batch size 1 and no gradient accumulation. Each scene is evaluated by accumulating softmax probabilities over its voxelized fragments before the final argmax. The test loader contains 312 ScanNet200 validation scenes; the other datasets use their official validation scenes under the same full-scene procedure. No test-time augmentation or query-label adaptation is used.

C.5 Calibration and Leakage Controls

DVD relation likelihoods, CBG margins, explanation priors, component-size thresholds, log-Bayes-factor thresholds, and boundary weights are calibrated on held-out base episodes. To emulate novel categories during calibration, subsets of base classes are temporarily designated as pseudo-novel, but their held-out labels are not used as query supervision. Once calibration is complete, all parameters are frozen. In particular, no novel query label, unsupported class name, or test-scene statistic is used to tune a threshold. Protected base points are excluded from OW-PC structural growth, whereas ambiguous base points remain exposed and may be revised only when a component-level explanation defeats the calibrated base and background counterfactuals.

C.6 Calibration Protocol Details

Calibration follows the same order as inference. Pseudo-novel episodes are formed by holding out a subset of base classes while retaining the remaining base classes as the protected vocabulary. The held-out classes provide positive proposal events for DVD, while the remaining classes provide negative proposal events and base/background counterfactuals. DVD relation frequencies are estimated with additive smoothing, the CBG score distribution is used to select τ_b, and the component priors, minimum sizes, and δ_h are fitted from held-out component statistics. Novel query labels and unsupported novel-class names are excluded from every step.

The calibration split is disjoint from the base scenes used to fit the segmentation backbone. Each calibrated quantity is stored with the feature-extractor and primary/auxiliary VLM identifiers, because changing either representation changes the scale of $G_i(\ell)$, support similarity, and the ddCRP feature term. A new VLM pairing therefore requires recalibration, while the structural inference and state definitions remain unchanged.

For partial-support episodes, the unknown state is calibrated from pseudo-novel components whose semantic support is intentionally removed. This makes the unknown prior and its minimum-size threshold applicable to withheld classes at evaluation time. In the zero-support condition, the same calibrated unknown state is used, while semantic matching and category counting are postponed until after component inference.

For reproducibility, each run records the number of pseudo-novel episodes, the held-out classes per episode, the smoothing constant δ, the selected τ_b, the state-specific $n_{\min}(h)$ and δ_h, and the

Table 7: Five-seed results for the full model under the 1-shot setting.

	ScanNet			ScanNet200		
Seed	Base	Novel	HM	Base	Novel	HM
42	78.44	51.92	62.48	68.51	34.80	46.16
43	78.20	52.17	62.58	67.93	34.74	45.97
123	79.15	50.14	61.39	69.31	35.40	46.86
44	78.46	49.37	60.61	67.66	35.52	45.58
45	79.01	52.29	62.93	67.87	34.78	45.99
Mean $\pm$ std	78.65 $\pm$ 0.41	51.18 $\pm$ 1.33	62.00 $\pm$ 0.97	68.26 $\pm$ 0.67	35.05 $\pm$ 0.38	46.31 $\pm$ 0.39
	ScanNet++			S3DIS		
Seed	Base	Novel	HM	Base	Novel	HM
42	64.84	23.25	34.23	60.35	33.85	43.18
43	64.97	23.15	34.14	58.80	35.52	44.16
123	63.47	22.52	33.25	60.16	38.05	46.60
44	64.09	24.10	35.03	59.29	35.23	44.14
45	63.59	22.92	33.70	58.18	32.64	41.76
Mean $\pm$ std	64.19 $\pm$ 0.69	23.19 $\pm$ 0.58	34.07 $\pm$ 0.66	59.36 $\pm$ 0.91	35.06 $\pm$ 2.03	43.97 $\pm$ 1.77

fraction of exposed elements entering structural inference. These quantities are diagnostics and are never tuned on novel validation scenes.

C.7 Metrics and Aggregation

For semantic segmentation, we compute per-class intersection-over-union and report the unweighted mean over the base classes (Base), novel classes (Novel), and all classes (All classes). The harmonic mean is $\mathrm{HM} = 2\,\mathrm{Base}\,\mathrm{Novel}/(\mathrm{Base}+\mathrm{Novel})$. In the partial-prototype setting we additionally report Novel with support and Novel without support; their class-count-weighted combination equals All novel, and the base/novel groups are then combined by class count to obtain All classes. All displayed values are percentages and are rounded only after aggregation.

For support-free recovery, each ground-truth novel instance is matched to the predicted unknown component with maximum IoU. Instance recall at IoU 0.50 is the fraction of ground-truth instances whose best match is at least 0.50, while mean best IoU averages the best match over all ground-truth instances, including missed ones. This matching is strictly post-hoc and does not change the predicted partition. Finally, the novel-category count in Table 2 uses the deduplicated decoding protocol described in Section A.5; it measures vocabulary coverage and is therefore distinct from instance recall and from component fragmentation.

D Further Experimental Analysis

This section consolidates the supplementary experiments already available and specifies the controlled comparisons that will be added. Unless stated otherwise, all thresholds are calibrated on held-out base episodes and then fixed before novel-class evaluation.

D.1 Multi-Seed Stability

Table 7 reports five-seed results for the full model under the 1-shot setting. Results for ScanNet, ScanNet200, ScanNet++, and S3DIS are shown; all values are percentages and all standard deviations are computed across seeds.

D.2 Recovery Beyond Pointwise VLM Predictions

We measure novel precision, novel recall, and missed-point recall on held-out 1-shot evaluations. Missed-point recall is computed only over ground-truth novel points that GFS-VL fails to predict as novel. This diagnostic directly tests whether the method merely filters existing predictions or recovers support-compatible query structure outside their pointwise coverage.

Table 8: Post-hoc semantic-bottleneck diagnostics (1-shot).

Dataset	Method	Novel P	Novel R	Missed R
ScanNet	GFS-VL	61.96	60.30	0.00
	Ours	68.48	64.76	16.40
ScanNet++	GFS-VL	43.34	53.24	0.00
	Ours	51.32	67.03	39.77
ScanNet200	GFS-VL	83.26	82.27	0.00
	Ours	83.54	82.99	23.46

Table 9: Class-wise recall on the ScanNet200 validation set (1-shot).

Method	mRecall-A	mRecall-B	mRecall-N	Micro Recall
GFS-VL	49.32	**65.60**	44.98	**81.95**
Ours	**58.83**	65.39	**57.09**	81.79
Δ	+9.51	-0.21	+12.10	-0.16

Table 10: Single-VLM and module diagnostics on 1-shot ScanNet and ScanNet++.

ScanNet 1-shot: single-VLM blind-spot diagnostic							
Method	mIoU-B	mIoU-N	HM	Novel P	Novel R	Missed-point recovery	Zero-coverage recovery ($\geq$25%)
Single-VLM baseline	78.56	48.93	60.30	61.96	60.30	0.00	0.00
DVD	79.31	53.45	63.86	68.48	64.76	16.40	6.25
ScanNet++ 1-shot: controlled module ablation							
Variant	mIoU-B	mIoU-N	HM	Novel P	Novel R	Missed-point recovery	Base$\rightarrow$novel leakage
Single-VLM baseline	55.98	26.20	35.70	38.39	72.47	0.00	15.52
DVD	58.03	26.95	36.80	41.63	69.91	11.90	12.55
DVD + CBG	61.32	26.09	36.60	42.78	63.77	13.00	10.07
DVD + CBG + OW-PC	**64.36**	**27.74**	**38.77**	**58.77**	63.33	**17.74**	**5.47**

Table 10 further reports a controlled module diagnostic. Novel P/R denote novel-point precision and recall. Missed-point recovery is recall over novel points missed by the single-VLM prediction; zero-coverage recovery is the fraction of previously uncovered novel instances reaching at least 25% correct coverage; and base-to-novel leakage is the fraction of base points mislabeled as novel; all metrics are percentages.

Table 8 (Novel P/R: novel-point precision/recall; Missed R: recall on novel points missed by GFS-VL; all values in %) shows that our method improves novel recall on all three datasets and recovers novel points missed by GFS-VL. In Table 10, adding OW-PC to DVD+CBG gives the highest missed-point recovery on ScanNet++, while CBG reduces base-to-novel leakage. Together, these measurements distinguish genuine recovery from indiscriminate novel expansion.

D.3 Detailed ScanNet200 Recall

Table 9 reports class-wise recall on the full ScanNet200 validation set under the 1-shot setting; A/B/N denote all 57 classes, the 12 base classes, and the 45 novel classes, Micro Recall is the point-level recall over all classes, and Δ is Ours minus GFS-VL. The improvement in novel-class mean recall, together with nearly unchanged base recall, is consistent with CBG exposing only ambiguous base predictions and OW-PC requiring marginalized component evidence before relabeling them.

D.4 ScanNet200 Partial-Prototype Class Breakdown

This additional evaluation partitions the main ScanNet200 split's 45 novel classes into 22 classes with semantic support prototypes and 23 classes without support prototypes; it does not change the main 12-base/45-novel definition. Table 11 lists the five best-performing classes in each group, averaged over the five registration checkpoints; its bottom row is the mean IoU over these five listed classes, not over the full group. The aggregate partial-prototype results are reported in the main paper. For the unsupported group, class names are used only for post-hoc evaluation and are not supplied to OW-PC during inference.

Table 11: Five best classes per group under the ScanNet200 partial-prototype split.

With support prototype		Without support prototype	
Class	IoU	Class	IoU
ceiling	77.89	bathtub	49.15
tv	69.33	plant	39.76
monitor	66.45	kitchen counter	39.68
couch	63.14	nightstand	29.06
coffee table	57.30	shower curtain	27.01
Top-5 mIoU	66.82	Top-5 mIoU	36.93

D.5 SUPPORT-AVAILABILITY DETAILS

Standard GFS evaluation assumes complete novel support, so it cannot test whether a method recovers novel instances without semantic supervision. We evaluate two additional regimes under the same base–novel partition and identical calibration: *partial support*, which keeps prototypes for a subset of novel classes and withholds them for the rest, and *zero support*, which supplies no novel text, prototype, or semantic logit and lets OW-PC return category-agnostic unknown instances that are matched to the ground truth only post-hoc.

Partial support. Table 5 shows that OW-PC improves every novel metric over GFS-VL. The largest gains are on the classes whose prototypes are withheld: novel-without-support mIoU rises from 7.05 to 37.02 on ScanNet, from 4.01 to 16.91 on ScanNet++, from 7.08 to 25.54 on ScanNet200, and from 42.89 to 46.64 on S3DIS. Unsupported classes therefore still yield usable structure, which we attribute to the unknown hypothesis: a component without named evidence is accepted as an unnamed instance instead of being forced into a supported class or background. Supported-class and base mIoU are preserved or improved, e.g., 78.99 versus 77.74 on ScanNet, so the gain does not trade off the known vocabulary, and All-novel and All-class mIoU increase on all four datasets. Table 11 lists the strongest classes in each ScanNet200 group.

Zero support. Table 4 shows that, with no novel semantics at all, OW-PC still recovers more novel instances than every baseline. On S3DIS it reaches 73.32% instance recall at IoU 0.50 versus 21.12% for GFS-VL, and 18.90%/15.56% for attMPTI/GW, together with 69.07% mean best IoU versus 31.34%. The gains are smaller but consistent on ScanNet, ScanNet++, and ScanNet200, at 25.52% versus 24.63%, 20.28% versus 17.54%, and 26.96% versus 26.01%, with mean best IoU improving by 2.7–4.3 points; the narrower ScanNet200 margin reflects its 45-class vocabulary, where a single unnamed component is harder to match to one ground-truth instance.

Hyperparameter stability. Figure 3 shows that the two structural thresholds are not sensitive: sweeping the CBG margin τ_b and the ddCRP margin δ_h over $\{0.02, 0.05, 0.10, 0.20\}$ on ScanNet200 1-shot changes Base, Novel, and HM by at most 0.05 points across all 16 combinations, and full-support accuracy is retained in Table 1.

The near-flat surface has a direct explanation: both margins are decision boundaries, so they can only move elements that sit close to those boundaries. An element enters the protected base set only when the primary and auxiliary views agree on the same base class and its base margin exceeds τ_b; for the large majority of base elements that margin lies far above the swept range, so changing τ_b leaves their status unchanged. Component acceptance behaves the same way: it compares a candidate explanation's log-Bayes factor against δ_h, and components with strong anchor or unknown-object evidence sit far above the threshold while clearly inconsistent merges sit far below it. Neither margin therefore reaches the bulk of the query. The elements already retained by the teacher, the elements both views accept with high confidence, and the elements covered by an explicit support prototype keep their labels wherever τ_b and δ_h are placed inside the swept interval. These are exactly the elements that dominate the supervision and hence most of the training objective, so the objective is largely independent of the two thresholds. Only the thin boundary population whose base margin or component log-Bayes factor falls inside the interval can switch, and it is this population that accounts for the residual 0.05-point variation. This is the intended consequence of assigning anchor precision, base stability, and scene-level recall to separate variables rather than trading them through

Table 12: Controlled primary–auxiliary VLM pairings on ScanNet200 under the 1-shot setting (seed 42). "Single" uses only the listed primary VLM with no auxiliary adjudication, while "DVD" combines the two models in the indicated roles. All values are percentages; the best entry in each metric column is in **bold**.

Method	Primary VLM	Auxiliary VLM	Base	Novel	HM
Single	OpenScene	—	64.38	24.99	34.24
Single	Mosaic	—	64.89	**31.16**	39.75
DVD	OpenScene	Mosaic	65.01	24.62	35.71
DVD	Mosaic	OpenScene	**66.03**	30.06	**41.31**

one dense pseudo-label threshold: no single margin governs a large fraction of the query, so no single sweep can move the aggregate metrics far.

D.6 Primary–Auxiliary VLM Pairings

Table 12 reports a controlled comparison on ScanNet200 under the 1-shot protocol (seed 42), covering two single-VLM baselines and both role assignments of the OpenScene–Mosaic pair. Two observations follow. First, the primary backbone dominates: the stronger Mosaic model reaches 39.75 HM and 31.16 novel mIoU on its own, well above the 34.24 and 24.99 of OpenScene, and each DVD configuration stays close to the accuracy of its own primary model. This is consistent with DVD using the primary VLM to generate candidate anchors and the auxiliary VLM only to adjudicate them, so proposal quality sets an upper bound that verification cannot fully recover. Second, the role assignment is not symmetric: with the same two encoders, the same DVD/CBG/OW-PC pipeline, and identical calibration, Mosaic-primary/OpenScene-auxiliary reaches 41.31 HM against 33.98 for the reversed assignment. The gap shows that which model proposes and which verifies is a real design variable, and it motivates comparing every pair in both directions rather than treating the two VLMs as interchangeable.

With the stronger model as the primary, DVD also improves on the best single-VLM baseline: HM rises from 39.75 to 41.31 and base mIoU from 64.89 to 66.03, while novel mIoU is essentially unchanged (30.06 versus 31.16). Placing the weaker OpenScene model as the primary reverses the effect and drops HM and novel mIoU below the single-VLM baselines, indicating that auxiliary adjudication helps only when the primary view already proposes reliable candidates. We therefore read this table as evidence that the designed role asymmetry matters, rather than as a final ranking of pairings.

The pairings also clarify why two VLMs are used at all. With a single VLM, base-class precision and novel-class recall are coupled through one dense threshold: making the novel evidence strict enough to keep base-class accuracy high also removes many correct novel predictions, whereas relaxing it lets base points be absorbed into novel anchors. The dual-VLM design decouples the two decisions. The primary VLM, together with the episode support, proposes novel candidates, and the auxiliary VLM adjudicates them from an independent representation, so an anchor is committed to only when the two views agree; this is what keeps base-class accuracy high. Rejected candidates are not deleted from the query: they remain in the CBG-exposed region and are passed to OW-PC, which fixes their labels from marginalized component evidence. The auxiliary view is therefore used to raise the precision of the anchors we trust, not to shrink the set of novel regions we consider, so better base-class precision is obtained without broadly discarding reasonable novel-class predictions. Table 12 is consistent with this reading: relative to the strongest single-VLM baseline, the dual-VLM configuration keeps base mIoU at 66.03 versus 64.89 while retaining almost all of the novel mIoU (30.06 versus 31.16), and the role asymmetry shows that this balance is reached only when the stronger model leads the proposal stage.

We will supplement the current evaluation with controlled pairings of different primary VLMs and auxiliary VLMs. The comparison will follow four rules. First, the primary VLM will be fixed while the auxiliary VLM is varied, isolating verification quality. Second, the auxiliary VLM will be fixed while the primary VLM is varied, isolating proposal and feature quality. Third, each compatible pair will be evaluated in both role assignments to test whether the designed asymmetry is empirically

necessary. Fourth, same-model primary/auxiliary controls will measure the loss of complementary evidence when the two views are strongly correlated.

For every pairing, we will report anchor precision and recall after DVD, protected-base precision and coverage after CBG, novel precision and recall, missed-point recall, base and novel mIoU, and HM after OW-PC. Calibration will be repeated only on held-out base episodes for each pairing; novel query labels will not be used to tune thresholds. Runtime and memory will also be reported because different encoders may change feature extraction cost even when the structural inference is unchanged.

D.7 Sensitivity, Support Quality, and Efficiency

The extended evaluation will vary the DVD anchor threshold, CBG margin, ddCRP spatial and feature scales, self-link mass, support-mode budget, and the state-specific component-size and LBF thresholds. We will additionally inject controlled corruption into labeled support points and auxiliary predictions. Each sweep will report both novel recovery and base leakage, since a recall increase without the corresponding precision and base-stability measurements cannot establish safe structural completion.

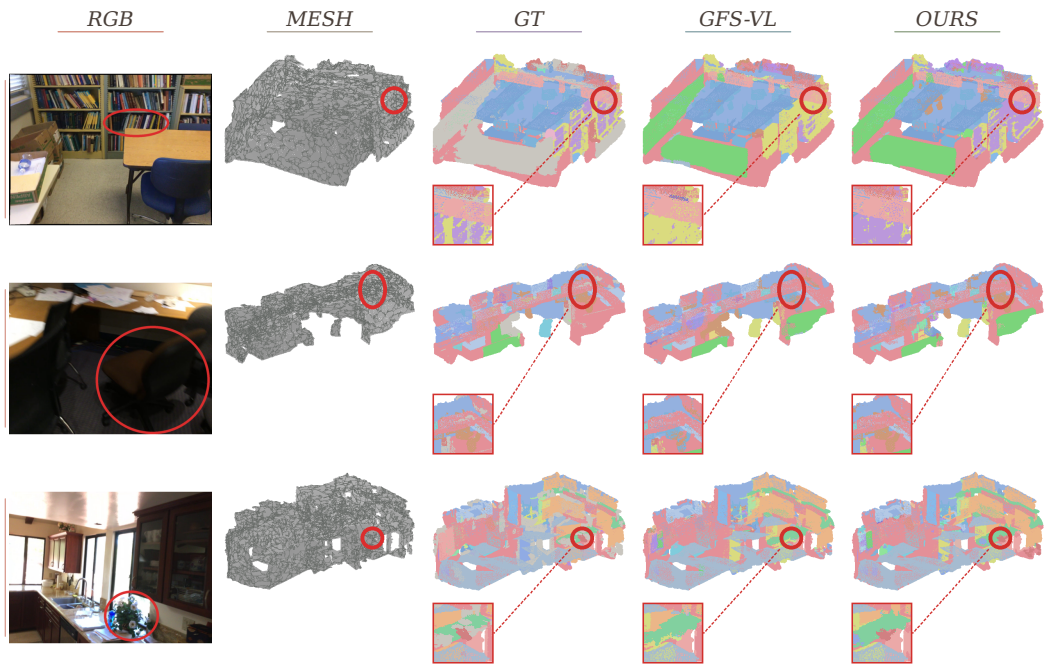
E More Qualitative Visualizations

The supplementary visualization will use a fixed camera, voxelization, and color assignment across all methods so that differences cannot be attributed to rendering. Each case will contain the raw query point cloud, the support-defined novel prototype, the primary and auxiliary predictions, noisy VLM novel candidates, DVD anchors, CBG-protected base points, the OW-PC component partition, the final prediction, and ground truth. Novel classes will retain one consistent accent color across stages; protected base classes will use fixed class-specific colors; exposed but unassigned structure will remain gray.

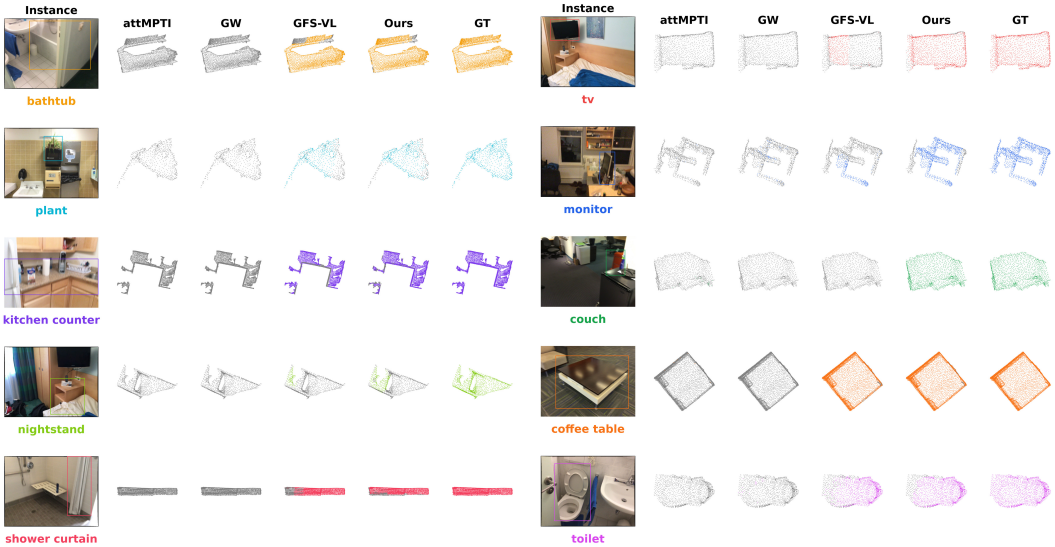
Figure 4a presents qualitative comparisons under the standard generalized few-shot setting, where labeled support prototypes are available for all novel classes.

Figure 4c shows novel-instance results under the partial-prototype setting, in which labeled prototypes are retained for 50% of the novel classes.

With half of the novel-class prototypes retained, OW-PC uses the available support evidence to complete object regions that remain fragmented or absent in the GFS-VL predictions. The gain is most pronounced for the monitor and couch, while the television and toilet predictions also cover more of the ground-truth extent. For the coffee table, both methods already recover nearly the complete instance, indicating that OW-PC preserves strong baseline predictions rather than expanding every component indiscriminately.



(a) Standard generalized few-shot setting. Columns show the RGB observation, reconstructed mesh, ground truth, GFS-VL prediction, and our prediction. Red circles and insets indicate the highlighted regions.



(b) Support-free novel classes. Columns show the instance crop, attMPTI, GW, GFS-VL, Ours, and ground truth. Bright colors indicate recalled target points, while gray indicates missed target points.

(c) Novel classes with 50% support prototypes. Columns show the instance crop, attMPTI, GW, GFS-VL, Ours, and ground truth. Bright colors indicate recalled target points, while gray indicates missed target points.

Figure 4: Qualitative comparisons under the standard generalized few-shot setting (a), on support-free novel classes (b), and on novel classes with 50% support prototypes (c). Bright colors indicate recalled target points, while gray indicates missed target points.