DENSITY-GUIDED VARIATIONAL SEMANTIC FIELD RECOVERY FOR OPEN-VOCABULARY 3D PART SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Point–text matching based methods for open-vocabulary 3D part segmentation directly convert prompt-conditioned responses into part predictions, making them vulnerable to low-density semantic outliers and boundary-mixed responses in the underlying semantic field. To address this issue, we formulate the task as prompt-conditioned semantic field recovery before final prediction and propose DGV-SFR. First, density-guided semantic denoising moves unreliable point states toward local prompt-consistent high-density modes while suppressing propagation across risky part boundaries. Second, a unified variational Dirichlet–Wiener correction jointly models the nonparametric posterior semantic surface and pointwise uncertainty, and evolves uncertain states on this surface under boundary and anchor constraints. The recovered semantic field is finally used for standard masked point–text matching, preserving the original open-vocabulary prediction interface. Experiments on standard benchmarks show that DGV-SFR achieves state-of-the-art average performance, improving a strong point–text matching baseline by approximately 4.1 mIoU. Further ablations, calibration analysis, and boundary-missing stress tests show that the recovered field reduces low-density errors, improves uncertainty calibration, and better preserves segmentation quality when local boundary evidence is incomplete.

1 INTRODUCTION

Open-vocabulary 3D part segmentation aims to assign fine-grained part labels to 3D points according to text prompts rather than a fixed taxonomy (Koo et al., 2022; Liu et al., 2023b; Zhou et al., 2023; Ma et al., 2025). This capability is important for interactive 3D perception, robotic manipulation, and shape editing, where the relevant part vocabulary varies across objects and user queries. Unlike closed-set segmentation, the challenge is therefore not only to recognize predefined geometric categories, but to maintain reliable point-level semantics under a language-defined and dynamically changing label space.

Recent methods typically address this problem by transferring language-aligned visual representations into 3D and matching point features with text embeddings (Radford et al., 2021; Zhang et al., 2022; Zhu et al., 2023; Liu et al., 2023a; Xue et al., 2023; 2024; Liu et al., 2023b; Zhou et al., 2023; Ma et al., 2025). Multi-view evidence, region proposals, prompt masks, and object-context prompts further improve this matching while preserving the open-vocabulary interface. However, point–text matching based methods largely assume that the prompt-conditioned semantic field before prediction is already re-

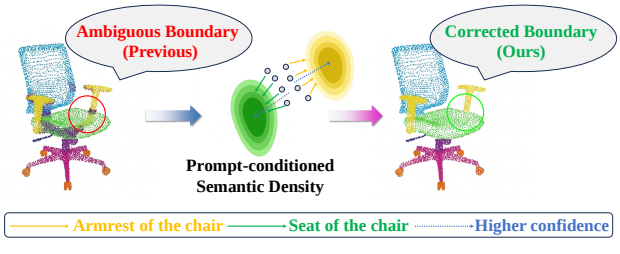

Figure 1: Prompt-conditioned semantic field recovery. Direct point–text matching can turn low-density outliers and mixed boundary responses into part-label errors. DGV-SFR recovers the field before matching by combining density-guided denoising with Dirichlet–Wiener correction, yielding cleaner prompt responses and sharper part boundaries.

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.

liable. For fine-grained 3D parts, this assumption is fragile. Sparse and incomplete observations can produce low-density semantic outliers that deviate from locally compatible semantic structures, while adjacent or geometrically similar parts often yield mixed prompt responses near boundaries. These unreliable semantic states are then directly inherited by the final point–text matching stage, leading to incorrect part predictions, as illustrated by the low-density outliers and mixed boundary responses in Fig. 1.

We address this reliability bottleneck by reformulating open-vocabulary 3D part segmentation as prompt-conditioned semantic field recovery before final prediction, and propose DGV-SFR, a layer-wise framework that progressively restores the intermediate semantic field. The framework contains two complementary recovery modules followed by the unchanged open-vocabulary prediction interface. Density-Guided Semantic Denoising first corrects low-density semantic outliers by moving unreliable point states toward locally compatible and prompt-consistent high-density modes while suppressing unsafe propagation across part boundaries. Variational Dirichlet–Wiener Correction then handles the residual uncertainty around ambiguous regions by modeling a posterior semantic surface and conservatively correcting uncertain point states under boundary constraints. Together, the two modules improve semantic field reliability while preserving the original open-vocabulary prediction interface.

Specifically, Density-Guided Semantic Denoising constructs local affinities from geometry, feature similarity, prompt-response consistency, and boundary risk, and performs reliability-gated mode seeking over the resulting semantic neighborhood. Building on the denoised field, Variational Dirichlet–Wiener Correction uses a truncated Dirichlet process to capture an adaptive set of latent semantic modes and induce a nonparametric posterior semantic surface, while Wiener dynamics evolve uncertain point states on this surface under boundary projection and anchor constraints. The corresponding variational posterior jointly models semantic modes, point-wise uncertainty, and correction trajectories, coupling semantic recovery and uncertainty estimation within a unified probabilistic process. The recovered field is finally fed into standard masked point–text matching for open-vocabulary prediction. As also visualized in Fig. 1, recovering the semantic field before matching produces cleaner prompt responses and sharper part boundaries, providing more reliable evidence for the final label decision. Experiments on standard benchmarks show that DGV-SFR achieves state-of-the-art average performance and improves a strong point–text matching baseline by approximately 4.1 mIoU.

Our contributions are as follows:

- We introduce Density-Guided Semantic Denoising, a local prompt-conditioned mode-seeking module that combines geometry, feature similarity, prompt agreement, and boundary barriers.
- We propose a unified Variational Dirichlet–Wiener Correction process whose joint variational posterior includes nonparametric semantic modes, uncertainty, and boundary-constrained Wiener correction paths.
- We provide main comparisons, targeted ablations, field-reliability analyses, calibration metrics, a boundary-missing stress test, and efficiency measurements that directly test the proposed recovery mechanism.

2 Related Work

Open-vocabulary 3D part segmentation. Open-vocabulary 3D understanding has advanced through language-image pretraining and 3D-language alignment. CLIP-style supervision (Radford et al., 2021) and 3D extensions such as PointCLIP (Zhang et al., 2022), PointCLIP V2 (Zhu et al., 2023), ULIP (Xue et al., 2023), ULIP-2 (Xue et al., 2024), OpenShape (Liu et al., 2023a), CLIP2Scene (Chen et al., 2023), and OpenScene (Peng et al., 2023) improve transferable 3D representations. Dense open-vocabulary 3D segmentation further uses point- or region-level matching, as in PLA (Ding et al., 2023), OpenMask3D (Takmaz et al., 2023), Open3DIS (Nguyen et al., 2024), RegionPLC (Yang et al., 2024), Search3D (Takmaz et al., 2024), and foundation-model-based 3D segmentation (Jiang et al., 2024). Prompt-conditioned part understanding extends this line to fine-grained object parts, including PartGlot (Koo et al., 2022), PartSLIP (Liu et al., 2023b), PartSLIP++ (Zhou et al., 2023), PartDistill (Umam et al., 2024), ZeroPS (Xue et al., 2025), PARIS3D (Kareem

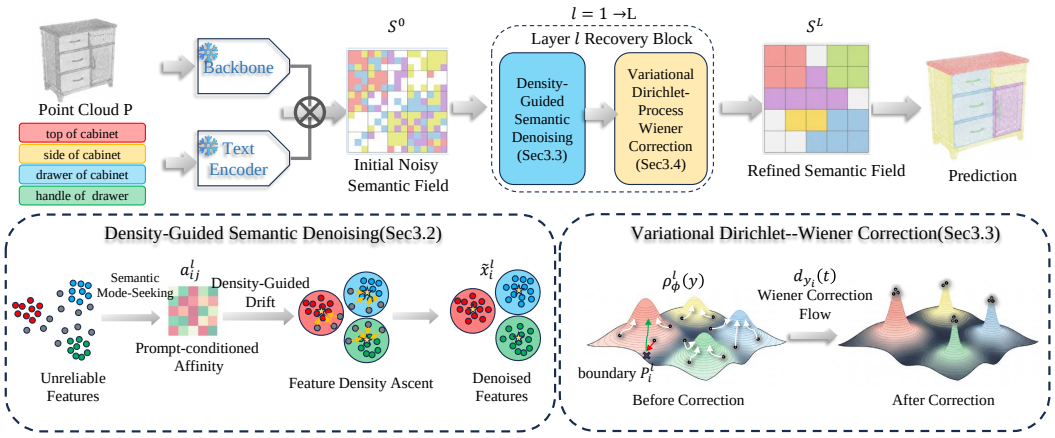

Figure 2: Overall framework of DGV-SFR. The model builds an initial prompt-conditioned semantic field, applies density-guided denoising to move unreliable points toward local high-density modes, and uses variational Dirichlet–Wiener correction to update uncertain states on a posterior semantic surface. The recovered field is passed to the same masked point–text matching rule, preserving the open-vocabulary interface.

et al., 2024), Reasoning3D (Chen et al., 2024), Find3D (Ma et al., 2025), and CoSMo3D (Jin et al., 2026). These methods improve 3D grounding and part prediction, but many still rely on final point–text matching. Our work keeps that interface and focuses on recovering the prompt-conditioned semantic field before matching.

Semantic denoising and mode-seeking refinement. Local refinement, graph neural networks, and attention-based point operators are common tools for dense 3D prediction (Qi et al., 2017; Wang et al., 2019b; Zhao et al., 2021; Wu et al., 2022; Wang et al., 2019a; 2018). They improve consistency by aggregating neighboring evidence, but fine-grained part segmentation requires a conservative aggregation rule because nearby points may belong to different parts. Density-based mode seeking provides a useful interpretation: points can be moved toward local high-density modes under a feature distribution (Fukunaga & Hostetler, 1975; Comaniciu & Meer, 2002). Related ideas appear in manifold regularization, graph-based learning, dense CRFs, contrastive compactness, and calibration-aware refinement (Belkin et al., 2006; Zhu et al., 2003; Kraehenbuehl & Koltun, 2011; Khosla et al., 2020; Wen et al., 2016; Guo et al., 2017). Our denoising module is prompt-conditioned and boundary-aware: affinity is defined jointly over geometry, feature state, text-response agreement, and boundary risk, so it is not ordinary geometric smoothing.

Variational nonparametric uncertainty and stochastic flows. Mixture models and Bayesian nonparametrics represent multi-modal data through latent components, while Dirichlet-process mixtures allow the number of active components to adapt to observations (Ferguson, 1973; Sethuraman, 1994; Neal, 2000; Blei & Jordan, 2006). Variational inference provides tractable posterior approximations for such latent-variable models (Jordan et al., 1999; Blei et al., 2017; Kingma & Welling, 2014). Separately, stochastic processes and score-driven dynamics move samples along learned probability surfaces (Ito, 1951; Oksendal, 2003; Hyvarinen, 2005; Song & Ermon, 2019; Song et al., 2021). For part segmentation, however, unconstrained diffusion can blur boundaries. We therefore couple posterior surface learning and Wiener correction in one variational process: the posterior provides density, score, and uncertainty, while boundary projection and anchoring make the correction conservative.

3 Method

3.1 Problem Formulation and Framework Overview

Problem formulation. Let $\mathcal{P} = \{p_i\}_{i=1}^N$ be an input point cloud, where $p_i \in \mathbb{R}^3$ is the coordinate of point i and N is the number of points. Let $\mathcal{T} = \{t_k\}_{k=1}^K$ be the text prompt set, where $t_k \in \mathbb{R}^d$ is the embedding of prompt k and K is the number of candidate part prompts for the current object.

A binary mask $M \in \{0,1\}^K$ marks valid prompts, with $M_k = 1$ if prompt k is valid. Let $x_i^\ell \in \mathbb{R}^d$ denote the point feature state at recovery layer ℓ, with $\ell = 0$ for the initial state and $\ell = L$ for the final recovered state. We use $\mathcal{S}^\ell = \{(x_i^\ell, r_i^\ell)\}_{i=1}^N$ for the prompt-conditioned semantic field at layer ℓ. Bars denote ℓ_2 normalization.

Direct matching forms a prompt-conditioned field by

$$s_{ik}^\ell = \tau \langle \bar{x}_i^\ell, \bar{t}_k \rangle, \qquad r_{ik}^\ell = \frac{M_k \exp(s_{ik}^\ell)}{\sum_{q:M_q=1} \exp(s_{iq}^\ell)}, \tag{1}$$

where s_{ik}^ℓ is the point–prompt score, r_{ik}^ℓ is the masked probability, and $\tau > 0$ is the temperature. The denominator contains only valid prompts, so invalid entries are zero and $r_i^\ell \in \Delta^{K-1}$; objects without valid prompts are excluded. Recovery therefore changes the semantic evidence without changing the prompt-defined label space.

Framework overview. The pipeline in Fig. 2 has two jointly optimized recovery modules and a fixed prediction interface. Density-Guided Semantic Denoising (Sec. 3.2) moves unreliable responses toward prompt-consistent modes while boundary barriers suppress cross-part propagation. Variational Dirichlet–Wiener Correction (Sec. 3.3) models residual multimodal uncertainty with a truncated Dirichlet process and evolves uncertain points on its posterior surface under projection and anchoring. Sec. 3.4 jointly supervises both modules without changing prompts, masks, or matching.

3.2 Density-Guided Semantic Denoising

The first module targets low-density semantic outliers. Reliable points of the same part tend to form compatible high-density neighborhoods, whereas noisy or ambiguous responses often lie on low-density outskirts. For point i, let $\mathcal{N}(i)$ be its k nearest neighbors. The boundary-aware affinity, local semantic mode, and denoised feature are

$$\begin{aligned} a_{ij}^\ell &= (1 - b_{ij}^\ell) \exp\left(-\frac{\|p_i - p_j\|_2^2}{\sigma_p^2} - \frac{\|x_i^\ell - x_j^\ell\|_2^2}{\sigma_f^2} - \frac{\mathrm{JS}(r_i^\ell, r_j^\ell)}{\sigma_s^2} \right), \\ m_i^\ell &= \begin{cases} \dfrac{\sum_{j\in\mathcal{N}(i)} a_{ij}^\ell x_j^\ell}{\sum_{j\in\mathcal{N}(i)} a_{ij}^\ell}, & \sum_{j\in\mathcal{N}(i)} a_{ij}^\ell > 0, \\ x_i^\ell, & \text{otherwise}, \end{cases} \qquad \tilde{x}_i^\ell = \mathrm{Norm}\big(x_i^\ell + \eta_i^\ell (m_i^\ell - x_i^\ell)\big). \end{aligned} \tag{2}$$

Here a_{ij}^ℓ is the affinity, $b_{ij}^\ell \in [0,1]$ the learned boundary barrier, and JS the Jensen–Shannon divergence of masked prompt distributions. Specifically, $b_{ij}^\ell = \mathrm{sigmoid}(\mathrm{MLP}_b([p_i - p_j, x_i^\ell - x_j^\ell, r_i^\ell - r_j^\ell]))$, where MLP_b outputs one scalar; larger values therefore suppress the pair. For the density interpretation below, write $a_{ij}^\ell = \bar{a}_{ij}^\ell \exp(-\|x_i^\ell - x_j^\ell\|_2^2/\sigma_f^2)$, where $\bar{a}_{ij}^\ell$ contains the coordinate, prompt-response, and boundary factors. Let $A_i^\ell = \sum_{j\in\mathcal{N}(i)} a_{ij}^\ell$, $H_i^\ell = -\sum_{k:M_k=1} r_{ik}^\ell \log r_{ik}^\ell$ with $0 \log 0 = 0$, and $\beta_i^\ell = \max_{j\in\mathcal{N}(i)} b_{ij}^\ell$. The reliability gate is $g_i^{\mathrm{rel},\ell} = \mathrm{sigmoid}(\mathrm{MLP}_\eta([A_i^\ell, H_i^\ell, \beta_i^\ell]))$ and $\eta_i^\ell = \eta g_i^{\mathrm{rel},\ell}$, so $\eta = 0.14$ is a global strength rather than a pointwise value. The bandwidths $\sigma_p, \sigma_f, \sigma_s$ correspond to coordinates, features, and prompt responses, and Norm is fixed ℓ_2 normalization. A zero-affinity neighborhood leaves the state unchanged; otherwise m_i^ℓ is a weighted local mode.

Holding the neighborhood and the coordinate, prompt-response, and boundary factors in $\bar{a}_{ij}^\ell$ fixed, the displacement $m_i^\ell - x_i^\ell$ in Eq. 2 is a mean-shift direction on a prompt-conditioned kernel density. For $A_i^\ell > 0$,

$$\hat{\rho}_i^\ell(x) = \sum_{j\in\mathcal{N}(i)} \bar{a}_{ij}^\ell \exp\left(-\frac{\|x - x_j^\ell\|_2^2}{\sigma_f^2} \right), \qquad m_i^\ell - x_i^\ell = \frac{\sigma_f^2}{2} \nabla_x \log \hat{\rho}_i^\ell(x)\big|_{x=x_i^\ell}, \tag{3}$$

which follows from the Gaussian-kernel mean-shift identity (Fukunaga & Hostetler, 1975; Comaniciu & Meer, 2002). The feature kernel appears once in both the gradient and the center in Eq. 2; the identity is local and does not guarantee part separation.

The interpolation cleans compatible neighborhoods while retaining high-risk states, leaving residual boundary ambiguity for the next module.

3.3 VARIATIONAL DIRICHLET–WIENER CORRECTION

After denoising, boundary-adjacent points may remain compatible with several modes. The second module is one Dirichlet–Wiener correction process: the truncated Dirichlet process defines the posterior semantic surface and Wiener dynamics (Ito, 1951; Oksendal, 2003) evolve uncertain states on that surface. They are specified jointly below.

For layer ℓ, let $y_i^{\ell,0} = \psi_\theta([\tilde{x}_i^\ell, r_i^\ell]) \in \mathbb{R}^h$ be the denoised state, where $[\cdot,\cdot]$ denotes concatenation, ψ_θ is a learned projection, and h is the semantic-state dimension. We treat this deterministic projection as the observed initial state and write $Y_i^\ell(0) = y_i^{\ell,0}$; it is not sampled a second time. With R maximum components, the truncated process is (Ferguson, 1973; Sethuraman, 1994; Neal, 2000)

$$\begin{aligned} v_m^\ell &\sim \mathrm{Beta}(1,\alpha), \quad \pi_m^\ell = v_m^\ell \prod_{r<m}(1-v_r^\ell), \quad m = 1,\dots,R-1, \\ \pi_R^\ell &= \prod_{r<R}(1-v_r^\ell), \quad \Theta_m^\ell = (\mu_m^\ell, \Sigma_m^\ell), \quad m = 1,\dots,R, \\ c_i^\ell &\sim \mathrm{Cat}(\pi^\ell), \quad Y_i^\ell(0) = y_i^{\ell,0}, \\ dY_i^\ell(t) &= f_\phi^\ell(Y_i^\ell(t), u_i^\ell, P_i^\ell)\,dt + \nu_i^\ell dW_i(t). \end{aligned} \tag{4}$$

Here v_m^ℓ is a stick variable, $\alpha > 0$ the concentration parameter, and π_m^ℓ the mixture weight; the final stick absorbs the remainder, so $\sum_m \pi_m^\ell = 1$. Θ_m^ℓ contains mean $\mu_m^\ell \in \mathbb{R}^h$ and covariance $\Sigma_m^\ell \in \mathbb{R}^{h\times h}$. The assignment c_i^ℓ selects a component; $Y_i^\ell(t)$ is its semantic path, $W_i(t)$ Wiener noise, P_i^ℓ the boundary projection, $u_i^\ell \in [0,1]$ uncertainty, and $\nu_i^\ell \geq 0$ its noise scale. R is a truncation bound, not the number of active modes.

The joint density used by the variational objective factors as

$$\begin{aligned} p(\mathcal{Y}^\ell, \mathcal{W}^\ell, \mathcal{C}^\ell, V^\ell, \Theta^\ell \mid \mathcal{T}) &= p_\alpha(V^\ell) p_0(\Theta^\ell) p(\mathcal{C}^\ell \mid V^\ell) \\ &\quad \times p(\mathcal{Y}^\ell \mid \mathcal{C}^\ell, \Theta^\ell) \times p(\mathcal{W}^\ell \mid \mathcal{Y}^\ell, \mathcal{C}^\ell, V^\ell, \Theta^\ell, u^\ell, P^\ell, \mathcal{T}). \end{aligned} \tag{5}$$

The factors in Eq. 5 are the stick, independent component-prior, assignment, Gaussian-likelihood, and Wiener-path factors. Here $p_0(\Theta^\ell) = \prod_m G_0(\Theta_m^\ell)$ is independent of V^ℓ, while the path conditions on V^ℓ through its posterior surface. The Gaussian term scores the observed $y_i^{\ell,0}$; it does not sample the initial state again. We use $q(v_m^\ell) = \mathrm{Beta}(a_m^\ell, b_m^\ell)$, $q(\mu_m^\ell, \Sigma_m^\ell) = \mathrm{NIW}(\bar{\mu}_m^\ell, \kappa_m^\ell, \Psi_m^\ell, \xi_m^\ell)$, and $q(c_i^\ell) = \mathrm{Cat}(\varphi_i^\ell)$. Here $a_m^\ell, b_m^\ell > 0$ are Beta shapes; the NIW arguments are posterior location, precision, scale, and degrees of freedom; and φ_{im}^ℓ are normalized assignments. Valid NIW parameters give $\Sigma_m^\ell \succ 0$. Euler training uses step Δt and noise $\nu_i^\ell \sqrt{\Delta t}\xi_i^t$, with $\xi_i^t \sim \mathcal{N}(0, I_h)$ and T training steps; inference uses $\nu_i^\ell = 0$ and one deterministic mean transition per block.

The drift used in Eq. 4 is defined on the posterior semantic surface:

$$\begin{aligned} \rho_\phi^\ell(y) &= \sum_{m=1}^R \mathbb{E}_{q_\phi^\ell}[\pi_m^\ell] \mathcal{N}\Big(y; \mathbb{E}_{q_\phi^\ell}[\mu_m^\ell], \mathbb{E}_{q_\phi^\ell}[\Sigma_m^\ell]\Big), \\ u_i^\ell &= \mathrm{Norm}_u\left(H(\varphi_i^\ell) + \sum_{m=1}^R \varphi_{im}^\ell \mathrm{tr}\Big(\mathbb{E}_{q_\phi^\ell}[\Sigma_m^\ell]\Big)\right), \\ f_\phi^\ell(Y_i, u_i, P_i) &= \gamma u_i^\ell P_i^\ell \nabla_y \log \rho_\phi^\ell(Y_i) - \lambda(1-u_i^\ell)(Y_i - y_i^{\ell,0}). \end{aligned} \tag{6}$$

The surface ρ_ϕ^ℓ is a plug-in posterior predictive density using expectations under q_ϕ^ℓ. Here $\varphi_{im}^\ell = q(c_i^\ell = m)$, H is assignment entropy, tr is matrix trace, and Norm_u maps nonnegative uncertainty to $[0,1]$. The positive γ and λ control score flow and anchoring.

The joint variational posterior explicitly includes the Wiener path:

$$q_\phi^\ell = \prod_{m=1}^{R-1} q(v_m^\ell) \prod_{m=1}^R q(\mu_m^\ell, \Sigma_m^\ell) \prod_{i=1}^N q(c_i^\ell) q\Big(Y_i^\ell(1{:}T) \mid Y_i^\ell(0) = y_i^{\ell,0}, c_i^\ell, u_i^\ell, P_i^\ell\Big), \tag{7}$$

where $Y_i^\ell(1{:}T)$ is the discretized path after the observed state. The mixture variables, uncertainty, and Wiener path are optimized jointly; this is not an independent diffusion appended to a DPMM. Training is stochastic, whereas inference is deterministic with $\nu_i^\ell = 0$.

The projection P_i^ℓ is an $h \times h$ positive semi-definite operator computed from the denoising neighborhood. Let $A_i^\ell = \sum_{j\in\mathcal{N}(i)} a_{ij}^\ell$ and define $\omega_{ij}^\ell = a_{ij}^\ell / A_i^\ell$ when $A_i^\ell > 0$. With these normalized compatibility weights, the boundary-weighted semantic covariance is

$$C_i^\ell = \sum_{j\in\mathcal{N}(i)} \omega_{ij}^\ell (1-b_{ij}^\ell)(y_j^{\ell,0}-y_i^{\ell,0})(y_j^{\ell,0}-y_i^{\ell,0})^\top, \qquad \omega_{ij}^\ell \geq 0, \quad \sum_{j\in\mathcal{N}(i)} \omega_{ij}^\ell = 1, \quad A_i^\ell > 0. \tag{8}$$

If $A_i^\ell = 0$, Eq. 8 is not evaluated: we set $C_i^\ell = 0$, $P_i^\ell = 0$, and bypass the correction update with $Y_i^\ell(t) = Y_i^\ell(0) = y_i^{\ell,0}$. Thus the zero-affinity branch has no undefined normalization or residual anchor drift. Otherwise, if $C_i^\ell \succeq 0$ and $\delta > 0$, then $P_i^\ell = C_i^\ell (C_i^\ell + \delta I_h)^{-1}$, with I_h the identity, suppresses risky score directions. For the local statement, omit anchor and noise terms and define the score-only trajectory by $\frac{dY_{i,\mathrm{score}}^\ell(t)}{dt} = \gamma u_i^\ell P_i^\ell \nabla_y \log \rho_\phi^\ell(Y_{i,\mathrm{score}}^\ell(t))$, with $Y_{i,\mathrm{score}}^\ell(0) = y_i^{\ell,0}$. If ρ_ϕ^ℓ is differentiable, then

$$\frac{d}{dt} \log \rho_\phi^\ell(Y_{i,\mathrm{score}}^\ell(t)) = \gamma u_i^\ell \left\| P_i^{\ell\,1/2} \nabla_y \log \rho_\phi^\ell(Y_{i,\mathrm{score}}^\ell(t)) \right\|_2^2 \geq 0. \tag{9}$$

The equality follows from the chain rule and positive semi-definiteness. It proves local posterior-density ascent only for the score component; the anchor can have either sign in the full drift, so boundary preservation remains an empirical effect of projection, anchoring, and conservative settings.

3.4 Training Objective and Prediction

The modules are optimized jointly. Let $\mathcal{Y}^\ell = \{y_i^{\ell,0}\}_{i=1}^N$, $\mathcal{C}^\ell = \{c_i^\ell\}_{i=1}^N$, $V^\ell = \{v_m^\ell\}_{m=1}^{R-1}$, $\Theta^\ell = \{\Theta_m^\ell\}_{m=1}^R$, and $\mathcal{W}^\ell = \{Y_i^\ell(1{:}T)\}_{i=1}^N$. The terminal state is $x_i^{\ell+1} = \mathrm{Norm}_x(Y_i^{\ell,T})$, where $\mathrm{Norm}_x : \mathbb{R}^h \to \mathbb{R}^d$ is a learned projection followed by fixed ℓ_2 normalization; $\mathcal{X}^L = \{x_i^L\}_{i=1}^N$ is the recovered field. The joint objective is

$$\begin{aligned} \max_{\phi,\theta} \sum_{\ell=0}^{L-1} \Big[\mathbb{E}_{q_\phi^\ell} \log p(\mathcal{Y}^\ell, \mathcal{W}^\ell, \mathcal{C}^\ell, V^\ell, \Theta^\ell \mid \mathcal{T}) - \mathbb{E}_{q_\phi^\ell} \log q_\phi^\ell \Big] \\ + \mathcal{L}_{\mathrm{seg}}(\mathcal{X}^L, \mathcal{T}; \theta) + \lambda_{\mathrm{bnd}} \mathcal{L}_{\mathrm{bnd}}(B; \theta). \end{aligned} \tag{10}$$

The bracket is the unit-weighted ELBO. Labels map to valid prompt indices g_i, giving $\mathcal{L}_{\mathrm{seg}} = N^{-1} \sum_i \log r_{i,g_i}^L$. Let $\mathcal{E}$ be the supervised neighbor-pair set, $B = \{b_{ij}^\ell\}$, and $y_{ij}^{\mathrm{bnd}} = \mathbb{1}[g_i \neq g_j]$. We define the boundary term as the average binary log-likelihood

$$\mathcal{L}_{\mathrm{bnd}} = \frac{1}{L|\mathcal{E}|} \sum_{\ell=0}^{L-1} \sum_{(i,j)\in\mathcal{E}} \left[y_{ij}^{\mathrm{bnd}} \log b_{ij}^\ell + (1 - y_{ij}^{\mathrm{bnd}}) \log(1 - b_{ij}^\ell) \right]. \tag{11}$$

It is non-positive, so the plus sign in Eq. 10 is correct for maximization; it is not a positive BCE loss. The coefficient λ_{bnd} is fixed after validation. Because paths and final states share semantic coordinates, prediction uses x_i^L in Eq. 1.

4 Experiments

4.1 Experimental Setup

The evaluation tests semantic field recovery without changing the prediction interface. Controlled variants share initial features, prompts, masks, and protocol; cross-method comparisons follow each method's reported protocol on 3DCoMPaT200-Coarse, 3DCoMPaT200-Fine, ShapeNetPart, and PartNetE (Ahmed et al., 2024; Yi et al., 2016; Mo et al., 2019). Literature rows align evaluation

Table 1: Main comparison under aligned open-vocabulary 3D part segmentation evaluation settings; literature rows retain their reported training protocols. Values are mIoU (%). Best and second-best results are in bold and underlined, respectively. "Obj-Part" denotes Part-of-Object prompts. Asterisks denote alternative protocol variants reported by prior work; the superscript † identifies the dataset-specific PartSLIP++ training variant reported for ShapeNetPart and PartNetE.

mIoU (%)	**Venue**	**3DCoMPaT200-Coarse**				**3DCoMPaT200-Fine**			
		Canonical		Rotated		Canonical		Rotated	
		Obj-Part	Part	Obj-Part	Part	Obj-Part	Part	Obj-Part	Part
PointCLIPV2	ICCV'23	14.09	14.16	13.17	13.39	7.25	7.18	7.43	7.20
PartSLIP	CVPR'23	1.27	6.54	1.74	8.03	0.23	1.56	0.71	4.44
PartSLIP++	arXiv'23	6.12	17.62	5.89	15.82	3.79	5.68	3.41	4.76
Find3D	ICCV'25	31.72	25.36	32.47	26.23	12.69	12.57	12.73	12.79
Find3D*	ICCV'25	36.89	45.00	38.14	46.02	17.72	24.97	18.57	26.35
CoSMo3D	CVPR'26	<u>47.51</u>	<u>54.52</u>	<u>47.74</u>	<u>54.55</u>	<u>24.48</u>	<u>31.29</u>	<u>24.46</u>	<u>30.97</u>
Ours	–	**52.36**	**59.28**	**52.08**	**59.64**	**29.18**	**35.27**	**28.92**	**35.81**
mIoU (%)	**Venue**	**ShapeNetPart**				**PartNetE**			
		Canonical		Rotated		Canonical		Rotated	
		Obj-Part	Part	Obj-Part	Part	Obj-Part	Part	Obj-Part	Part
OpenMask3D	NeurIPS'23	8.94	10.37	6.75	14.56	12.54	11.24	11.93	11.67
PartSLIP++†	arXiv'23	1.43	6.46	0.94	6.03	5.12	**32.71**	3.87	**23.03**
PointCLIPV2	ICCV'23	16.91	20.22	16.88	18.19	11.28	9.70	10.32	10.22
Find3D	ICCV'25	28.39	24.09	29.64	23.71	16.86	16.38	17.62	17.16
CoSMo3D	CVPR'26	<u>36.16</u>	<u>33.31</u>	<u>34.20</u>	<u>32.84</u>	<u>17.59</u>	17.19	<u>18.48</u>	18.17
Ours	–	**38.08**	**37.86**	**36.32**	**37.58**	**20.74**	<u>21.62</u>	**22.04**	<u>22.41</u>

dimensions but need not share training data or feature extractors. The internal direct-matching baseline supports controlled ablations; CoSMo3D, Find3D variants, and dataset-specific PartSLIP++ retain their reported protocols. The asterisk and dagger mark the Find3D and PartSLIP++ variants.

The Part setting uses only the part name, whereas Part-of-Object adds object context; we report mIoU for both. Unless stated otherwise, DGV-SFR uses four recovery blocks, $k = 12$, $R = 16$, one Wiener step per block, AdamW (Loshchilov & Hutter, 2019), deterministic mean dynamics, $\eta = 0.14$, and $\gamma = 0.18$. At inference, $T_{\text{inf}} = 1$ and $\nu_i^\ell = 0$. During training, T is the number of Euler steps, Δt is their discretization step, ν_i^ℓ is the uncertainty-scaled noise coefficient, λ is the anchor strength, λ_{bnd} weights Eq. 11, $\alpha > 0$ is the Dirichlet concentration parameter, and $\delta > 0$ is the projection ridge. These quantities are selected once on the validation split and then held fixed for testing; their exact numerical defaults and sweep ranges must be recorded in the implementation configuration rather than inferred from the reported sensitivity plot. The ELBO has unit weight.

The boundary-missing stress test corrupts only the test input under canonical Part-of-Object prompts. For each object, 8% of points are removed around semantic boundaries; every method uses the same observed cloud, and nearest-neighbor lifting returns predictions to the original point set before mIoU is computed against complete ground truth. Training data, prompts, and model parameters remain unchanged.

Means are unweighted averages of the displayed prompt/pose cells. Rows marked by $*$ or † are separate protocol variants. The 32.72 reference is the internal shared-feature direct-matching variant in Table 3, not the externally reported CoSMo3D mean despite the same rounded value.

4.2 Main Results

DGV-SFR raises the shared-feature direct-matching baseline from 32.72 to 36.82 mean mIoU ($+4.10$). Among the reported literature rows, it has the highest displayed average and exceeds CoSMo3D by 4.76, 4.50, 3.33, and 3.85 points across the four groups; these margins are not a controlled retraining comparison. PartSLIP++† remains stronger on some PartNetE Part-prompt cells under dataset-specific supervision; Fig. 3 shows the corresponding boundary improvement.

The improvement persists across prompt constructions and poses, while the candidate label set remains unchanged. Coarse gains exceed fine gains because weaker fine-grained separation leaves

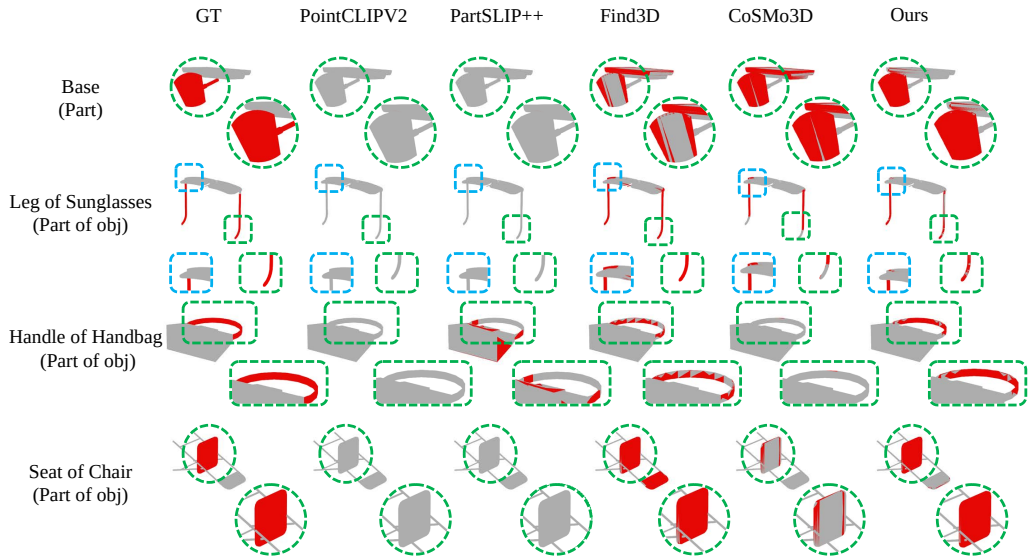

Figure 3: Qualitative comparison under the same prompt set. The examples compare direct point–text matching, DGV-SFR, and ground truth, illustrating how semantic field recovery removes isolated responses, sharpens boundaries, and keeps high-confidence regions stable.

Table 2: Boundary-missing stress test under canonical Part-of-Object prompts. The columns labeled 3DC-C, 3DC-F, SNP, and PNE denote 3DCoMPaT200-Coarse, 3DCoMPaT200-Fine, ShapeNet-Part, and PartNetE, respectively. Each dataset entry reports Hole mIoU (%) after removing 8% of points around semantic part boundaries at test time and evaluating on the original full point cloud after nearest-neighbor recovery. Avg. hole is the arithmetic mean over the four datasets. Bold and underlined values indicate the best and second-best Hole mIoU, respectively.

Method	3DC-C	3DC-F	SNP	PNE	Avg. hole
PointCLIPV2	31.54	6.46	23.06	14.21	18.82
PartSLIP++	1.06	0.37	16.91	1.12	4.87
Find3D	30.68	12.34	22.81	16.45	20.57
CoSMo3D	<u>46.46</u>	<u>23.31</u>	<u>34.38</u>	<u>17.94</u>	<u>30.52</u>
Ours	**51.38**	**28.14**	**37.11**	**20.46**	**34.27**

less local evidence to recover; the method cannot create information absent from the point cloud or prompt set. The selected PartSLIP++ advantage reflects dataset-specific supervision rather than the shared-feature comparison. Fig. 3 shows fewer isolated responses while retaining coherent high-confidence regions.

4.3 Boundary-Missing Stress Test

Under canonical Part-of-Object prompts, the stress test removes 8% of boundary points at inference and evaluates all methods after identical nearest-neighbor lifting (Table 2).

DGV-SFR leads on all four datasets (34.27 average Hole mIoU versus 30.52). Because lifting restores support but not deleted boundary evidence, the gain measures transfer under missing cues: denoising suppresses isolated responses and boundary projection limits cross-part motion. Smaller gains on weakly separated datasets indicate reliance on surviving neighborhoods.

Evaluation against complete ground truth tests whether assignments remain plausible when boundary points are unavailable. Identical nearest-neighbor lifting only returns predictions to common support. The advantage on every dataset suggests that the effect is not benchmark-specific, while the smaller margins on 3DCoMPaT200-Fine and PartNetE show that a posterior cannot reconstruct a thin or adjacent boundary when little evidence survives. The test therefore supports robustness to partial boundary evidence, not arbitrary deletion recovery.

4.4 Ablation Studies and Field Reliability

Ablations in Tables 3–5 remove one mechanism at a time while keeping features, prompts, protocol, and aggregation fixed.

Table 3: Density-guided denoising ablation. Mean mIoU (%) over the benchmark; all other modules are fixed.

Variant	Mean
Direct point–text matching	32.72
Generic MLP refinement	33.07
Density-guided denoising only	34.69
Full model	**36.82**

Table 4: Posterior surface and Wiener correction ablation. Mean mIoU (%) with one component changed.

Variant	Mean
Finite mixture semantic surface	34.22
Dirichlet-process mixture without Wiener correction	34.67
Wiener correction on raw graph affinity	34.01
Joint correction without ELBO supervision	35.36
Full model	**36.82**

Table 5: Conservative correction ablation. Mean mIoU (%) with one constraint removed.

Variant	Mean
w/o prompt-conditioned density	35.82
w/o boundary projection	36.25
w/o anchor tethering	36.22
Full model	**36.82**

Appendix reliability metrics show 9.4-point lower low-density error, 9.1-point lower boundary leakage, and improved calibration. The MLP baseline, surface/flow/ELBO variants, and constraint rows jointly show that density gating, posterior-surface learning, and conservative motion each contribute.

The denoising ablation separates mode seeking from generic capacity: the small MLP gain versus the larger denoising-only gain indicates that the initial correction targets low-density field errors. The posterior ablation tests the unified probabilistic stage, with each removal weakening adaptive surface modeling or path correction. Constraint results show that density, projection, and anchoring regulate where correction acts, its direction, and its magnitude.

4.5 Sensitivity, Efficiency, and Failure Modes

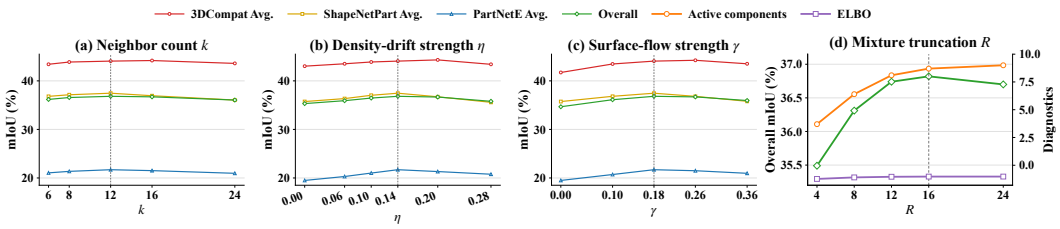

Figure 4: Hyperparameter sensitivity for neighbor count k, density-drift strength η, surface-flow strength γ, and mixture truncation R, corresponding to denoising neighborhood size, mode-seeking strength, Wiener score-flow strength, and mixture truncation level, respectively. Detailed sweep values, including the R-specific diagnostics, are reported in Appendix Table A6.

The selected settings ($k = 12$, $R = 16$, $\eta = 0.14$, $\gamma = 0.18$) lie in a stable local region: the 1.18M module raises mean mIoU from 32.72 to 36.82 at $1.31\times$ normalized time and $1.24\times$ normalized memory. Across the tested ranges, performance varies smoothly around these values rather than depending on a single isolated setting. The selected region is moderate; thin, weakly separated parts remain failures.

5 Conclusion

DGV-SFR reframes open-vocabulary 3D part segmentation as semantic field recovery before matching. Density-guided denoising removes low-density outliers, and Dirichlet–Wiener correction resolves residual uncertainty on a constrained posterior surface while preserving the prompt interface. The results support this view; thin or weakly separated parts, especially under severe partial observations, remain a limitation requiring stronger boundary evidence.

AI USE STATEMENT

In preparing this manuscript, generative AI tools were used to assist with English-language editing, organization of the manuscript, and consistency checks across the method description, equations, terminology, tables, and figure references. These tools were not used to generate or alter experimental data, fabricate results, or replace author verification of mathematical derivations and empirical claims. The authors reviewed all AI-assisted text and take responsibility for the final manuscript, including its equations, tables, figures, references, and reported conclusions.

ETHICS STATEMENT

This work studies open-vocabulary 3D part segmentation using existing public benchmark data and does not involve human subjects, private personal data, or a new data-collection process. The method is evaluated as a research system rather than as a safety-critical deployment. As with other 3D perception systems, inaccurate predictions caused by sparse observations, ambiguous prompts, or weak part boundaries could affect downstream robotic or design applications if used without validation. We therefore report boundary-missing stress tests, representative failure modes, and limitations, and recommend task-specific validation before deployment.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide the complete prompt-conditioned semantic field recovery formulation, notation, objective functions, model components, and inference procedure. They also specify the benchmark protocol, canonical and rotated evaluation settings, prompt construction, boundary-missing stress-test protocol, main hyperparameters, ablation variants, sensitivity ranges, and efficiency measurement procedure. The default recovery settings used in the reported experiments are $k = 12$, $R = 16$, $\eta = 0.14$, and $\gamma = 0.18$. The appendix further includes derivations, pseudocode for both recovery stages and the full inference pipeline, supplementary quantitative analyses, and implementation-level details needed to reproduce the reported comparisons under the same public benchmark protocol. All methods use the same prompt interface and evaluation aggregation unless explicitly stated otherwise.

REFERENCES

Mahmoud Ahmed, Xiang Li, Arpit Prajapati, and Mohamed Elhoseiny. 3DCoMPaT200: Language grounded large-scale 3D vision dataset for compositional recognition. In *NeurIPS*, 2024.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 2006.

David M. Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 2006.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *JASA*, 2017.

Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP. In *CVPR*, 2023.

Tianrun Chen, Chunan Yu, Jing Li, Jianqi Zhang, Lanyun Zhu, Deyi Ji, Yong Zhang, Ying Zang, Zejian Li, and Lingyun Sun. Reasoning3D: Grounding and reasoning in 3D: Fine-grained zero-shot open-vocabulary 3D reasoning part segmentation via large vision-language models. *arXiv preprint arXiv:2405.19326*, 2024.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE TPAMI*, 2002.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-driven open-vocabulary 3D scene understanding. In *CVPR*, 2023.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1973.

Keinosuke Fukunaga and Larry D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE TIT*, 1975.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Aapo Hyvarinen. Estimation of non-normalized statistical models by score matching. *JMLR*, 2005.

Kiyosi Ito. On stochastic differential equations. *Memoirs AMS*, 1951.

Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3D semantic segmentation with foundation models. In *CVPR*, 2024.

Li Jin, Weikai Chen, Yujie Wang, Yingda Yin, Zeyu Hu, Runze Zhang, Keyang Luo, Shengju Qian, Xin Wang, and Xueying Qin. CoSMo3D: Open-world promptable 3D semantic segmentation through LLM-guided canonical spatial modeling. In *CVPR*, 2026.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 1999.

Amrin Kareem, Jean Lahoud, and Hisham Cholakkal. PARIS3D: Reasoning-based 3D part segmentation using large multimodal model. In *ECCV*, 2024.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J. Guibas, and Minhyuk Sung. PartGlot: Learning shape part segmentation from language reference games. In *CVPR*, 2022.

Philipp Kraehenbuehl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NeurIPS*, 2011.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *NeurIPS*, 2023a.

Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. PartSLIP: Low-shot part segmentation for 3D point clouds via pretrained image-language models. In *CVPR*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Ziqi Ma, Yisong Yue, and Georgia Gkioxari. Find any part in 3D. In *ICCV*, 2025.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *CVPR*, 2019.

Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *JCGS*, 2000.

Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. In *CVPR*, 2024.

Bernt Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6 edition, 2003.

Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, 2023.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 1994.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Ayca Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *NeurIPS*, 2023.

Ayca Takmaz, Alexandros Delitzas, Robert W. Sumner, Francis Engelmann, Johanna Wald, and Federico Tombari. Search3D: Hierarchical open-vocabulary 3D segmentation. *arXiv preprint arXiv:2409.18431*, 2024.

Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. PartDistill: 3D shape part segmentation by vision-language model distillation. In *CVPR*, 2024.

Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *CVPR*, 2019a.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic Graph CNN for learning on point clouds. *ACM TOG*, 2019b.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer V2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martin-Martin, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In *CVPR*, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martin-Martin, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP-2: Towards scalable multimodal pre-training for 3D understanding. In *CVPR*, 2024.

Yuheng Xue, Nenglun Chen, Jun Liu, and Wenyun Sun. ZeroPS: High-quality cross-modal knowledge transfer for zero-shot 3D part segmentation. In *3DV*, 2025.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. In *CVPR*, 2024.

Li Yi, Vladimir G. Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas J. Guibas. A scalable active framework for region annotation in 3D shape collections. *ACM TOG*, 2016.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In *CVPR*, 2022.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021.

Yuchen Zhou, Jiayuan Gu, Xuanlin Li, Minghua Liu, Yunhao Fang, and Hao Su. PartSLIP++: Enhancing low-shot 3D part segmentation via multi-view instance segmentation and maximum likelihood estimation. *arXiv preprint arXiv:2312.03015*, 2023.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning. In *ICCV*, 2023.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. Technical report, Carnegie Mellon University, 2003.

APPENDIX CONTENTS

A Overview of the Supplementary Material 15
B Notation and Problem Setup 15
C Theoretical Details 15
C.1 Density-Guided Semantic Denoising as Local Mode Seeking 15
C.2 Unified Dirichlet–Wiener Generative Process 15
C.3 Posterior Surface, Uncertainty, and Drift 17
C.4 Joint Variational Posterior and Objective 17
C.5 Boundary Projection and Local Ascent 18
D Algorithms 19
E Experimental Protocol and Baseline Alignment 19
E.1 Prompt and Pose Protocols 19
E.2 Boundary-Missing Stress Test 19
F Supplementary Quantitative Analyses 21
F.1 Extended Boundary-Missing Robustness 21
F.2 Field-Reliability Metrics 21
F.3 Core Multi-Seed Stability 21
F.4 Hyperparameter Sensitivity 22
F.5 Efficiency and Complexity 23
G Qualitative Visualizations 25
H Limitations and Failure Modes 27

A Overview of the Supplementary Material

This supplementary material provides the technical and experimental details that support the main paper without changing the problem formulation, model interface, or evaluation protocol. Appendix B summarizes notation used by the recovery formulation. Appendix C expands the density-guided denoising step, the unified Dirichlet–Wiener correction process, and the variational objective. Appendix D gives implementation-level pseudocode for the two recovery modules and the full inference pipeline. Appendix E specifies prompt construction, pose settings, boundary-missing evaluation, and baseline alignment. Appendix F reports supplementary quantitative results for mechanism ablations, field reliability, prompt sensitivity, multi-seed stability, hyperparameter sweeps, and computational cost. Appendix G presents supplementary point-cloud, semantic-field, boundary-missing, and failure-case visualizations, and Appendix H records limitations consistent with the main-paper conclusion.

B Notation and Problem Setup

Problem formulation. The main paper defines open-vocabulary 3D part segmentation as prompt-conditioned semantic field recovery. Table A1 collects the symbols used in the main text and the supplementary derivations. Superscript ℓ denotes a recovery layer, subscript i denotes a point, subscript j denotes a neighboring point, subscript k denotes a text prompt, and subscript m denotes a latent semantic component.

C Theoretical Details

C.1 Density-Guided Semantic Denoising as Local Mode Seeking

Derivation. The density-guided denoising module in Eq. 2 is intended to correct unreliable semantic states before final prompt matching. For a fixed local neighborhood $\mathcal{N}(i)$ and fixed coordinate, prompt-response, and boundary factors, write $a_{ij}^{\ell}(x) = \bar{a}_{ij}^{\ell} K_j(x)$, where $\bar{a}_{ij}^{\ell}$ contains those non-feature factors. Define the prompt-conditioned kernel density around point i as

$$\hat{\rho}_i^{\ell}(x) = \sum_{j \in \mathcal{N}(i)} \bar{a}_{ij}^{\ell} \exp\left(-\frac{\|x - x_j^{\ell}\|_2^2}{\sigma_f^2} \right), \tag{12}$$

where $x \in \mathbb{R}^d$ is a candidate semantic state and the denoising affinity is $a_{ij}^{\ell} = a_{ij}^{\ell}(x_i^{\ell})$. Let

$$K_j(x) = \exp\left(-\frac{\|x - x_j^{\ell}\|_2^2}{\sigma_f^2} \right). \tag{13}$$

Differentiating Eq. 12 gives

$$\nabla_x \log \hat{\rho}_i^{\ell}(x) = \frac{2}{\sigma_f^2} \left(\frac{\sum_{j \in \mathcal{N}(i)} \bar{a}_{ij}^{\ell} K_j(x) x_j^{\ell}}{\sum_{j \in \mathcal{N}(i)} \bar{a}_{ij}^{\ell} K_j(x)} - x \right). \tag{14}$$

Interpretation. Evaluating the identity at $x = x_i^{\ell}$ gives exactly the center m_i^{ℓ} in Eq. 2; the feature kernel is counted once, not twice. This identifies the direction of the update, not the normalized and gated update itself, and does not assume global clustering or guaranteed part separability.

C.2 Unified Dirichlet–Wiener Generative Process

Generative construction. The second recovery module is a single probabilistic correction process. The truncated Dirichlet process determines a posterior semantic surface, and the Wiener correction path is defined on that surface. For layer ℓ, the denoised state is projected to

$$y_i^{\ell,0} = \psi_\theta\left([\tilde{x}_i^{\ell}, r_i^{\ell}] \right), \tag{15}$$

Table A1: Notation used by the prompt-conditioned semantic field recovery formulation. The table expands the symbols introduced in the main paper and keeps the same layer, point, prompt, and component indexing conventions.

Symbol	Meaning
$\mathcal{P} = \{p_i\}_{i=1}^{N}$	Input point cloud with N points; $p_i \in \mathbb{R}^3$ is the coordinate of point i.
$\mathcal{T} = \{t_k\}_{k=1}^{K}$	Object-specific text prompt set with K candidate part prompts; $t_k \in \mathbb{R}^d$ is the prompt embedding.
$M \in \{0,1\}^K$	Prompt validity mask; $M_k = 1$ means prompt k is valid for the current object taxonomy.
x_i^ℓ, $\bar{x}_i^\ell$	Point feature state at recovery layer ℓ and its ℓ_2-normalized version.
$\bar{t}_k$	ℓ_2-normalized text embedding of prompt k.
s_{ik}^ℓ, r_{ik}^ℓ	Masked point–text score and masked prompt probability from Eq. 1.
τ	Matching temperature used in normalized probabilities.
$\mathcal{N}(i)$, k	Local neighborhood of point i and the number of nearest neighbors.
a_{ij}^ℓ, $\bar{a}_{ij}^\ell$, $K_j(x)$, b_{ij}^ℓ	Final prompt-conditioned affinity, its non-feature base factor, feature kernel, and learned boundary barrier between point i and neighbor j.
$\sigma_p, \sigma_f, \sigma_s$	Bandwidths for coordinate distance, feature-state distance, and prompt-response divergence.
μ_i^ℓ, η_i^ℓ	Local compatible semantic mode and reliability-gated denoising strength.
$\tilde{x}_i^\ell$	Denoised feature state after boundary-aware mode seeking.
$y_i^{\ell,0}$	Projected semantic state initialized from $[\tilde{x}_i^\ell, r_i^\ell]$ before Dirichlet–Wiener correction.
ψ_θ, h	Learned projection into an h-dimensional semantic-state space.
R, m	Truncation level and component index of the variational Dirichlet process.
v_m^ℓ $(m < R)$, $v_R^\ell = 1$, α	Non-final stick variables, deterministic remainder stick, and concentration parameter.
π_m^ℓ	Mixture weight induced by the stick-breaking construction.
$\Theta_m^\ell = (\mu_m^\ell, \Sigma_m^\ell)$	Mean and covariance of latent semantic component m.
c_i^ℓ, φ_{im}^ℓ	Latent component assignment and its variational responsibility $q(c_i^\ell = m)$.
$\rho_\phi^\ell(y)$	Posterior predictive semantic surface used by the correction drift.
$\tilde{\pi}_m^\ell$, $\tilde{\mu}_m^\ell$, $\tilde{\Sigma}_m^\ell$	Plug-in posterior weight, mean, and stabilized covariance used to evaluate ρ_ϕ^ℓ.
u_i^ℓ	Posterior uncertainty of point i, computed from assignment entropy and component dispersion.
$Y_i(t)$, $W_i(t)$	Continuous correction path and standard Wiener process.
P_i^ℓ	Positive semi-definite boundary projection that suppresses high-risk correction directions.
C_i^ℓ, ω_{ij}^ℓ	Boundary-weighted semantic covariance and normalized neighborhood compatibility weight.
I_h, ε_Σ	h-dimensional identity and small covariance ridge used for numerical stabilization.
γ, λ	Surface-flow strength and anchor-tethering strength in the correction drift.
ν_i^ℓ, Δt, ξ_i^t	Training-time stochasticity scale, Euler step size, and standard Gaussian innovation.
$\mathcal{L}_{\text{seg}}$, $\mathcal{L}_{\text{bnd}}$	Masked point–text segmentation loss and boundary supervision loss.

where $[\cdot,\cdot]$ denotes concatenation. The component weights are generated by a finite stick-breaking construction. The last component absorbs the remaining mass:

$$v_m^\ell \sim \mathrm{Beta}(1,\alpha), \qquad \pi_m^\ell = v_m^\ell \prod_{r<m}(1-v_r^\ell), \qquad m = 1,\dots,R-1, \tag{16}$$

$$v_R^\ell = 1, \qquad \pi_R^\ell = \prod_{r=1}^{R-1}(1-v_r^\ell). \tag{17}$$

Thus $\sum_{m=1}^R \pi_m^\ell = 1$ for every realization of the non-final sticks. Each component has parameters $\Theta_m^\ell = (\mu_m^\ell, \Sigma_m^\ell)$ drawn from a base distribution G_0. For each point, the observed state contributes the likelihood

$$c_i^\ell \sim \mathrm{Cat}(\pi_1^\ell,\dots,\pi_R^\ell), \qquad p(y_i^{\ell,0} \mid c_i^\ell = m) = \mathcal{N}(y_i^{\ell,0}; \mu_m^\ell, \Sigma_m^\ell). \tag{18}$$

The deterministic value $Y_i^\ell(0) = y_i^{\ell,0}$ is initialized once; this likelihood scores the observed state and does not resample the path initial condition. The correction path then follows

$$dY_i(t) = f_\phi^\ell(Y_i(t), u_i^\ell, P_i^\ell)dt + \nu_i^\ell dW_i(t), \qquad Y_i(0) = y_i^{\ell,0}. \tag{19}$$

Coupling rationale. This definition makes the Wiener dynamics conditional on the semantic surface induced by the variational Dirichlet posterior. It is therefore not an independent graph diffusion step applied after a mixture model.

C.3 Posterior Surface, Uncertainty, and Drift

Posterior construction. The variational responsibilities $\varphi_{im}^\ell = q(c_i^\ell = m)$ define the posterior predictive surface. We use a plug-in approximation to the exact posterior expectation of the mixture density:

$$\begin{aligned} \tilde\pi_m^\ell &= \mathbb{E}_{q_\phi^\ell}[\pi_m^\ell], \qquad \tilde\mu_m^\ell = \mathbb{E}_{q_\phi^\ell}[\mu_m^\ell], \\ \tilde\Sigma_m^\ell &= \mathbb{E}_{q_\phi^\ell}[\Sigma_m^\ell] + \varepsilon_\Sigma I_h, \\ \rho_\phi^\ell(y) &\approx \sum_{m=1}^R \tilde\pi_m^\ell \mathcal{N}\left(y; \tilde\mu_m^\ell, \tilde\Sigma_m^\ell\right), \qquad \varepsilon_\Sigma > 0. \end{aligned} \tag{20}$$

The approximation replaces an expectation of a Gaussian mixture by a mixture evaluated at expected parameters; it is not an exact marginalization of q_ϕ^ℓ. In implementation, the covariance ridge is used only to keep Cholesky factorization well-conditioned, and mixture log weights are normalized with log-sum-exp before evaluating the score. The ridge is chosen below the feature-state scale and does not define an additional semantic component. **Uncertainty decomposition.** Point uncertainty combines assignment ambiguity and component dispersion:

$$u_i^\ell = \mathrm{Norm}_u\left(H(\varphi_i^\ell) + \sum_{m=1}^R \varphi_{im}^\ell \mathrm{tr}\left(\mathbb{E}_{q_\phi}[\Sigma_m^\ell]\right)\right), \tag{21}$$

where $H(\varphi_i^\ell) = -\sum_m \varphi_{im}^\ell \log \varphi_{im}^\ell$ is assignment entropy and Norm_u maps the raw score to $[0,1]$. The drift used by the discretized correction is

$$f_\phi^\ell(Y_i, u_i, P_i) = \gamma u_i^\ell P_i^\ell \nabla_y \log \rho_\phi^\ell(Y_i) - \lambda(1-u_i^\ell)(Y_i - y_i^{\ell,0}). \tag{22}$$

Drift interpretation. The first term moves uncertain points toward higher posterior semantic density, and the second term keeps confident points close to their denoised states. During inference, the stochastic scale is set to zero and Eq. 19 is evaluated by its deterministic mean dynamics.

C.4 Joint Variational Posterior and Objective

Variational factorization. Let $\mathcal{Y}^\ell = \{y_i^{\ell,0}\}_{i=1}^N$, $\mathcal{C}^\ell = \{c_i^\ell\}_{i=1}^N$, $V^\ell = \{v_m^\ell\}_{m=1}^{R-1}$, $\Theta^\ell = \{\Theta_m^\ell\}_{m=1}^R$, and $\mathcal{W}^\ell = \{Y_i^\ell(1{:}T)\}_{i=1}^N$. The deterministic initial states $Y_i^\ell(0) = y_i^{\ell,0}$ are conditioned on, not

included among the path random variables. The approximate posterior includes stick-breaking variables, component parameters, assignments, and correction paths:

$$q_\phi^\ell = \prod_{m=1}^{R-1} q(v_m^\ell) \prod_{m=1}^{R} q(\mu_m^\ell, \Sigma_m^\ell) \prod_{i=1}^{N} q(c_i^\ell) q\Big(Y_i^\ell(1{:}T) \mid Y_i^\ell(0) = y_i^{\ell,0}, c_i^\ell, u_i^\ell, P_i^\ell\Big), \tag{23}$$

where the deterministic value $v_R^\ell = 1$ is not a variational random variable. The complete joint factorization is

$$\begin{aligned} &p(\mathcal{Y}^\ell, \mathcal{W}^\ell, \mathcal{C}^\ell, V^\ell, \Theta^\ell \mid \mathcal{T}) \\ &\quad = p_\alpha(V^\ell) p_0(\Theta^\ell) p(\mathcal{C}^\ell \mid V^\ell) p(\mathcal{Y}^\ell \mid \mathcal{C}^\ell, \Theta^\ell) \\ &\qquad \times p(\mathcal{W}^\ell \mid \mathcal{Y}^\ell, \mathcal{C}^\ell, V^\ell, \Theta^\ell, u^\ell, P^\ell, \mathcal{T}). \end{aligned} \tag{24}$$

Here $p_0(\Theta^\ell) = \prod_m G_0(\Theta_m^\ell)$, $p(\mathcal{C}^\ell \mid V^\ell) = \prod_i \mathrm{Cat}(\pi^\ell)$, and $p(\mathcal{Y}^\ell \mid \mathcal{C}^\ell, \Theta^\ell) = \prod_i \mathcal{N}(y_i^{\ell,0}; \mu_{c_i^\ell}^\ell, \Sigma_{c_i^\ell}^\ell)$. With Euler step size Δt, the path factor is represented by

$$Y_i^{\ell,t+1} = Y_i^{\ell,t} + \Delta t\, f_\phi^\ell(Y_i^{\ell,t}, u_i^\ell, P_i^\ell) + \nu_i^\ell \sqrt{\Delta t}\, \xi_i^t, \qquad \xi_i^t \sim \mathcal{N}(0, I_h). \tag{25}$$

For $\nu_i^\ell > 0$, the corresponding discrete path likelihood is

$$p(\mathcal{W}^\ell \mid \mathcal{Y}^\ell, \cdots) = \prod_{i=1}^{N} \prod_{t=0}^{T-1} \mathcal{N}\Big(Y_i^{\ell,t+1}; Y_i^{\ell,t} + \Delta t f_\phi^\ell(Y_i^{\ell,t}, u_i^\ell, P_i^\ell), (\nu_i^\ell)^2 \Delta t I_h\Big). \tag{26}$$

The notation $\cdots$ denotes the conditioning variables displayed in Eq. 24. When $\nu_i^\ell = 0$, this Gaussian transition is replaced by its deterministic point-mass limit, and inference uses the mean update rather than Wiener sampling. The variational factors use $q(v_m^\ell) = \mathrm{Beta}(a_m^\ell, b_m^\ell)$ for $m < R$, $q(v_R^\ell) = \delta_1$, $q(\mu_m^\ell, \Sigma_m^\ell) = \mathrm{NIW}(\bar{\mu}_m^\ell, \kappa_m^\ell, \Psi_m^\ell, \xi_m^\ell)$, and $q(c_i^\ell) = \mathrm{Cat}(\varphi_i^\ell)$. Here $a_m^\ell, b_m^\ell > 0$ are Beta shapes, δ_1 is a point mass at one, the NIW arguments are posterior location, precision, scale, and degrees of freedom, and φ_{im}^ℓ are normalized assignments. The path factor is the product of the Gaussian transitions in Eq. 26 during stochastic training and a point mass at the deterministic Euler trajectory when $\nu_i^\ell = 0$.

ELBO definition. With the same notation, the layerwise evidence lower bound is

$$\mathcal{L}_{\mathrm{ELBO}}^\ell = \mathbb{E}_{q_\phi^\ell} \log p(\mathcal{Y}^\ell, \mathcal{W}^\ell, \mathcal{C}^\ell, V^\ell, \Theta^\ell \mid \mathcal{T}) - \mathbb{E}_{q_\phi^\ell} \log q_\phi^\ell. \tag{27}$$

The training objective in Eq. 10 adds the masked point–text segmentation term and boundary supervision. This objective keeps the open-vocabulary prediction interface intact: prompts define the candidate labels, and the recovery model modifies the semantic field used by the final matching rule.

For completeness, let g_i be the valid prompt index of point i, let $\mathcal{E}$ be the supervised neighbor-pair set, let $B = \{b_{ij}^\ell\}$ collect the boundary barriers, and let $y_{ij}^{\mathrm{bnd}} = \mathbb{1}[g_i \neq g_j]$. The two auxiliary terms used by the joint objective are

$$\mathcal{L}_{\mathrm{seg}} = \frac{1}{N} \sum_{i=1}^{N} \log r_{i,g_i}^L, \tag{28}$$

$$\mathcal{L}_{\mathrm{bnd}} = \frac{1}{L|\mathcal{E}|} \sum_{\ell=0}^{L-1} \sum_{(i,j)\in\mathcal{E}} \left[y_{ij}^{\mathrm{bnd}} \log b_{ij}^\ell + (1 - y_{ij}^{\mathrm{bnd}}) \log(1 - b_{ij}^\ell) \right]. \tag{29}$$

The boundary term is a log-likelihood and is therefore non-positive under maximization; it is not a sign-reversed BCE hidden in the notation.

Bound justification. Let $z^\ell = (\mathcal{W}^\ell, \mathcal{C}^\ell, V^\ell, \Theta^\ell)$ collect the latent variables in the correction process. The variational term is a lower bound on the marginal log evidence:

$$\log p(\mathcal{Y}^\ell \mid \mathcal{T}) = \log \int q_\phi^\ell(z^\ell) \frac{p(\mathcal{Y}^\ell, z^\ell \mid \mathcal{T})}{q_\phi^\ell(z^\ell)}\, dz^\ell \tag{30}$$

$$\geq \mathbb{E}_{q_\phi^\ell} \left[\log p(\mathcal{Y}^\ell, z^\ell \mid \mathcal{T}) - \log q_\phi^\ell(z^\ell) \right] = \mathcal{L}_{\mathrm{ELBO}}^\ell. \tag{31}$$

Proof. The inequality follows from Jensen's inequality applied to the logarithm, and the final equality is the definition in Eq. 27. Thus, the objective optimizes a tractable variational lower bound rather than assuming exact posterior inference.

C.5 Boundary Projection and Local Ascent

Projection construction. The projection matrix P_i^ℓ suppresses directions that are likely to cross semantic boundaries. One implementation estimates a boundary-weighted covariance

$$C_i^\ell = \sum_{j \in \mathcal{N}_b(i)} (1 - b_{ij}^\ell)\omega_{ij}^\ell (y_j^{\ell,0} - y_i^{\ell,0})(y_j^{\ell,0} - y_i^{\ell,0})^\top, \tag{32}$$

where $\mathcal{N}_b(i)$ is the boundary neighborhood and ω_{ij}^ℓ is a nonnegative normalized compatibility weight. For the following property, assume $b_{ij}^\ell \in [0,1]$, $\omega_{ij}^\ell \geq 0$, $y_i^{\ell,0} \in \mathbb{R}^h$, $C_i^\ell = C_i^{\ell\top}$, and $\delta > 0$. If the neighborhood affinity is zero, the implementation sets $C_i^\ell = P_i^\ell = 0$ and bypasses the correction update. Otherwise, a stable positive semi-definite projection is

$$P_i^\ell = C_i^\ell (C_i^\ell + \delta I_h)^{-1}, \qquad \delta > 0. \tag{33}$$

Projection property. Under these assumptions, Eq. 32 is a sum of positive semi-definite outer products, so $C_i^\ell \succeq 0$. Consequently, if $\lambda_r \geq 0$ is an eigenvalue of C_i^ℓ, the corresponding eigenvalue of P_i^ℓ is $\lambda_r/(\lambda_r + \delta) \in [0,1)$. Hence P_i^ℓ is symmetric, positive semi-definite, and non-expansive in the spectral norm. **Proof.** Since $C_i^\ell = C_i^{\ell\top} \succeq 0$, write $C_i^\ell = U \operatorname{diag}(\lambda_r) U^\top$ with an orthogonal U. The same eigensystem diagonalizes $C_i^\ell + \delta I_h$, so

$$P_i^\ell = U \operatorname{diag}\left(\frac{\lambda_r}{\lambda_r + \delta}\right) U^\top.$$

All eigenvalues lie in $[0,1)$, which proves symmetry, positive semi-definiteness, and $\|P_i^\ell\|_2 < 1$. **Proposition (local posterior ascent).** Assume that ρ_ϕ^ℓ is differentiable, $\gamma \geq 0$, $u_i^\ell \in [0,1]$, and $P_i^\ell = P_i^{\ell\top} \succeq 0$. The deterministic score component of Eq. 22 locally ascends the posterior surface. Let $g_i(t) = \nabla_y \log \rho_\phi^\ell(Y_i(t))$ and consider $dY_i(t)/dt = \gamma u_i^\ell P_i^\ell g_i(t)$. **Proof.** Along this score-only trajectory, the chain rule gives

$$\frac{d}{dt} \log \rho_\phi^\ell(Y_i(t)) = g_i(t)^\top \frac{dY_i(t)}{dt} = \gamma u_i^\ell g_i(t)^\top P_i^\ell g_i(t) \tag{34}$$

$$= \gamma u_i^\ell \left\| P_i^{\ell\,1/2} g_i(t) \right\|_2^2 \geq 0. \tag{35}$$

Scope. This proposition explains the correction direction but does not state a global segmentation guarantee. Boundary preservation still depends on the learned boundary barrier, the projection, anchor tethering, and the conservative hyperparameters evaluated in the experiments.

D Algorithms

Algorithms A1–A3 summarize the implementation in a way that mirrors the method sections in the main paper. The notation follows Table A1.

E Experimental Protocol and Baseline Alignment

E.1 Prompt and Pose Protocols

The experiments use two prompt modes. The *Part* mode uses only the part name, such as *leg* or *handle*. The *Part-of-Object* mode includes object context, such as *chair leg* or *cabinet handle*. Invalid prompts are masked before final point–text matching for all variants. The canonical pose setting uses the benchmark orientation, while the rotated setting follows the benchmark pose-variation protocol. The proposed recovery module is evaluated as a semantic-field recovery method rather than a rotation-invariance module.

E.2 Boundary-Missing Stress Test

The boundary-missing stress test removes points around semantic part boundaries only at inference time. Training data, learned parameters, prompt sets, and prompt masks remain unchanged. Each

Algorithm A1 Density-Guided Semantic Denoising

Input: $\mathcal{P} \in \mathbb{R}^{N\times 3}$, $X^\ell \in \mathbb{R}^{N\times d}$, $R^\ell \in \mathbb{R}^{N\times K}$, valid-prompt mask M, neighbor count k
Output: $\tilde{X}^\ell \in \mathbb{R}^{N\times d}$
for each point i **do**
  Build the k-nearest-neighbor set $\mathcal{N}(i)$ from $\mathcal{P}$, or reuse the shared local graph
  for each neighbor $j \in \mathcal{N}(i)$ **do**
    Estimate $b_{ij}^\ell \in [0,1]$ from $[p_i - p_j, x_i^\ell - x_j^\ell, r_i^\ell - r_j^\ell]$
    Compute the base factor $\bar{a}_{ij}^\ell$ from coordinate distance, prompt-response divergence, and b_{ij}^ℓ, excluding the feature kernel
    Compute $K_j(x_i^\ell) = \exp(-\|x_i^\ell - x_j^\ell\|_2^2/\sigma_f^2)$
    Set $a_{ij}^\ell = \bar{a}_{ij}^\ell K_j(x_i^\ell)$ as the final affinity in Eq. 2
  end for
  if $A_i^\ell = \sum_{j\in\mathcal{N}(i)} a_{ij}^\ell > 0$ **then**
    Normalize affinities and compute $m_i^\ell = A_i^{\ell-1} \sum_j a_{ij}^\ell x_j^\ell$
    Predict $\eta_i^\ell = \eta g_i^{\mathrm{rel},\ell}$ from density, entropy, and boundary risk
    Update $\tilde{x}_i^\ell = \mathrm{Norm}(x_i^\ell + \eta_i^\ell(m_i^\ell - x_i^\ell))$
  else
    Set $m_i^\ell = x_i^\ell$ and $\tilde{x}_i^\ell = x_i^\ell$
  end if
end for
return $\tilde{X}^\ell$

Algorithm A2 Variational Dirichlet–Wiener Correction

Input: $\tilde{X}^\ell \in \mathbb{R}^{N\times d}$, $R^\ell \in \mathbb{R}^{N\times K}$, mask M, truncation level R, step size Δt, steps T
Output: corrected states $Y^{\ell,T} \in \mathbb{R}^{N\times h}$
Project denoised states to $y_i^{\ell,0} = \psi_\theta([\tilde{x}_i^\ell, r_i^\ell])$
Infer $q(v_m^\ell)$ for $m < R$, $q(\mu_m^\ell, \Sigma_m^\ell)$ for $m \le R$, and φ_{im}^ℓ
Set $v_R^\ell = 1$ and construct ρ_ϕ^ℓ with the plug-in rule in Eq. 20
for each point i **do**
  Compute uncertainty u_i^ℓ from assignment entropy and posterior component dispersion
  Estimate projection P_i^ℓ from boundary-weighted local semantic covariance
  Initialize $Y_i^{\ell,0} = y_i^{\ell,0}$
  for $t = 0$ to $T-1$ **do**
    Evaluate score $\nabla_y \log \rho_\phi^\ell(Y_i^{\ell,t})$
    Set $Y_i^{\ell,t+1} = Y_i^{\ell,t} + \Delta t f_\phi^\ell(Y_i^{\ell,t}, u_i^\ell, P_i^\ell)$
    if training with $\nu_i^\ell > 0$ **then**
      Add $\nu_i^\ell \sqrt{\Delta t}\, \xi_i^t$, where $\xi_i^t \sim \mathcal{N}(0, I_h)$
    end if
  end for
end for
return $Y^{\ell,T}$

method predicts labels on the partially observed point cloud, and predictions are lifted back to the original full point set using the same nearest-neighbor recovery rule before computing mIoU against the full ground truth. The main paper reports the 8% setting, while Table A2 provides the measured 12% and 16% extensions under the same full-recovery evaluator.

Protocol alignment. The comparison follows a common masked point–text evaluation contract. PointCLIPV2, PartSLIP, PartSLIP++, OpenMask3D, and Find3D are retained as external open-vocabulary baselines under the prompt and pose settings reported for the corresponding benchmark comparison. PartSLIP++ carries the dagger marker in the main table because that row uses the dataset-specific training variant on ShapeNetPart and PartNetE. Find3D and its starred row are kept distinct: the star denotes the alternative protocol reported by the prior work and is not averaged with the original Find3D entry. CoSMo3D is used to define the aligned evaluation protocol, including Part and Part-of-Object prompts together with canonical and rotated poses, but it is treated only as

Algorithm A3 Inference Pipeline for Open-Vocabulary 3D Part Segmentation

Input: $\mathcal{P}$, prompt embeddings $\mathcal{T}$, valid-prompt mask M, initial features X^0
Output: point-wise labels $\hat{y} \in \{1, \ldots, K\}^N$
for $\ell = 0$ to $L - 1$ **do**
    Compute masked prompt probabilities r_i^ℓ using Eq. 1
    Apply Algorithm A1 to obtain $\tilde{X}^\ell$
    Apply Algorithm A2 to obtain corrected states $Y^{\ell,T}$
    Decode $x_i^{\ell+1} = \mathrm{Norm}_x(Y_i^{\ell,T})$
end for
Compute final masked prompt probabilities from X^L and $\mathcal{T}$
Set $\hat{y}_i = \arg\max_{k:M_k=1} r_{ik}^L$ for each point i
return $\hat{y}$

a strong comparison point. DGV-SFR uses the same valid-prompt mask and final masked point–text matching rule. Only the internal shared-feature ablations isolate the effect of semantic-field recovery; external literature rows need not share the training data or feature extractor.

F Supplementary Quantitative Analyses

F.1 Extended Boundary-Missing Robustness

The main paper reports the 8% boundary-missing setting. Table A2 reports the measured higher-severity 12% and 16% extensions using the same canonical Part-of-Object prompt protocol, boundary-centered point-removal rule, and nearest-neighbor full-point recovery as the main-paper stress test. The formal comparison is restricted to CoSMo3D and DGV-SFR because the supplied external-baseline sanity runs use adapted preprocessing and evaluator alignment.

Table A2: Measured higher-severity boundary-missing results. The columns 3DC-C and 3DC-F denote 3DCoMPaT200-Coarse and 3DCoMPaT200-Fine, respectively; SNP and PNE denote ShapeNetPart and PartNetE. Each value is the supplied benchmark-level Hole mIoU (%) aggregate for the stated missing ratio; the table does not expand the canonical/rotated cells. The formal comparison uses CoSMo3D and DGV-SFR under the same full-recovery evaluator.

Method	Missing ratio	3DC-C	3DC-F	SNP	PNE	Avg. hole
CoSMo3D	12%	45.53	22.89	32.49	17.69	29.65
Ours	12%	**50.38**	**27.04**	**35.06**	**20.85**	**33.33**
CoSMo3D	16%	44.61	21.97	31.35	17.28	28.80
Ours	16%	**49.29**	**26.22**	**34.20**	**20.58**	**32.57**

The external PointCLIPV2, Find3D, and PartSLIP++ values are retained in the experiment report as adapted sanity checks, but are not inserted into this formal table because their preprocessing and evaluator alignment differ from the CoSMo3D–DGV-SFR comparison. Under the aligned formal protocol, the higher deletion rate lowers both methods' Hole mIoU, while DGV-SFR retains an aggregate margin of 3.68 points at 12% removal and 3.77 points at 16% removal. The canonical/rotated breakdown is not claimed by this compact aggregate table; the evaluation protocol itself remains unchanged from the main-paper stress test.

F.2 Field-Reliability Metrics

The main paper summarizes field-reliability trends because the body length is limited. Table A3 defines the metrics, and Table A4 reports the measured comparison among direct matching, density-guided denoising only, and the complete Dirichlet–Wiener correction. The intermediate column isolates the contribution of density-guided denoising before posterior-surface correction.

F.3 Core Multi-Seed Stability

Table A5 reports a compact stability analysis for the core mechanism comparisons. The setting is the same benchmark-level mean mIoU used by the main-paper ablations, with shared frozen features

Table A3: Definitions of semantic-field reliability metrics used to interpret the recovery mechanism. These metrics supplement mIoU and are not used as training-time target metrics unless stated otherwise.

Metric	Definition and interpretation
Low-density error	Error rate on points whose initial semantic states lie in low prompt-conditioned density regions.
High-density error	Error rate on points whose initial semantic states lie near high-density semantic modes.
Posterior density gain	Change in posterior predictive density after correction, measured on uncertain or boundary-adjacent subsets.
ECE	Expected calibration error of masked prompt probabilities. Lower values indicate better probability calibration.
NLL	Negative log-likelihood of the valid prompt label under the masked prompt distribution.
Brier score	Mean squared probability error for the valid prompt distribution.
Boundary leakage	Fraction of boundary-adjacent updates that move semantic states toward an incompatible neighboring part.
Boundary mIoU	mIoU restricted to boundary-adjacent points.

Table A4: Measured semantic-field reliability, calibration, and boundary-preservation summary. Direct denotes masked point–text matching without recovery, Denoising only applies the first recovery module, and Full corresponds to the complete DGV-SFR pipeline. Lower is better for error, ECE, NLL, Brier score, and leakage; higher is better for density gain, error–uncertainty correlation, and boundary mIoU.

Metric	Direct	Denoising only	Full	Direct-to-Full change
Low-density error (%)	47.2	40.6	37.8	-9.4
High-density error (%)	17.3	16.4	15.9	-1.4
Uncertain-point density gain	0.00	+0.29	+0.46	+0.46
Boundary-adjacent density gain	0.00	+0.18	+0.38	+0.38
ECE (%)	9.4	7.2	5.9	-3.5
NLL	0.921	0.846	0.793	-0.128
Brier score	0.196	0.172	0.154	-0.042
Error–uncertainty correlation	0.38	0.50	0.59	+0.21
Boundary leakage (%)	20.6	15.0	11.5	-9.1
Boundary mIoU (%)	29.8	31.4	32.8	+3.0

and prompt masks fixed across variants. Only stochastic training factors such as recovery-head initialization and data order vary across the three seeds.

For seed-level measurements $m_1, \ldots, m_S$, the reported mean is $\bar{m} = S^{-1} \sum_{s=1}^{S} m_s$ and the reported sample standard deviation is $s_m = \sqrt{(S-1)^{-1} \sum_{s=1}^{S} (m_s - \bar{m})^2}$, with $S = 3$ for the displayed seeds. We report the standard deviation rather than a confidence interval because this compact analysis contains only three seeds.

Table A5: Three-seed stability of mechanism-critical variants. Values are mean mIoU (%) $\pm$ standard deviation over seeds 17, 42, and 73; the gap is measured relative to the full-model mean.

Variant	Mean mIoU $\pm$ std.	Gap to full model
Direct point–text matching	32.72 ± 0.14	-4.10
Density-guided denoising only	34.69 ± 0.16	-2.13
Dirichlet-process mixture without Wiener correction	34.67 ± 0.19	-2.15
Wiener correction on raw graph affinity	34.01 ± 0.13	-2.81
Full model	$\mathbf{36.82 \pm 0.18}$	**0.00**

F.4 Hyperparameter Sensitivity

The main paper displays the most important trends for neighbor count k, density-drift strength η, and surface-flow strength γ. Table A6 provides the corresponding numeric sweep and includes the truncation-level sweep for R, which is referenced in the main-paper caption of Fig. 4. Each

sweep changes one parameter while keeping the remaining default settings fixed. For each benchmark, we first average the four displayed prompt/pose cells; 3DCoMPaT Avg. is the mean of the coarse and fine benchmark means, and Overall is the unweighted mean of the resulting means for 3DCoMPaT200-Coarse, 3DCoMPaT200-Fine, ShapeNetPart, and PartNetE.

Table A6: Measured hyperparameter sensitivity. 3DCoMPaT Avg. averages 3DCoMPaT200-Coarse and 3DCoMPaT200-Fine; Overall is the arithmetic mean of the four benchmark means. Active components and normalized ELBO are reported only for the truncation sweep because they are specific to R.

Parameter	Value	3DCoMPaT Avg.	ShapeNetPart Avg.	PartNetE Avg.	Overall	Active comps.	ELBO
k	6	43.29	36.95	21.40	36.23	–	–
k	8	43.67	37.18	21.82	36.59	–	–
k	12	44.07	37.46	21.70	36.82	–	–
k	16	43.92	37.30	21.55	36.67	–	–
k	24	43.43	36.97	20.82	36.16	–	–
η	0.00	43.20	36.72	20.80	35.98	–	–
η	0.06	43.57	37.04	21.18	36.34	–	–
η	0.10	43.84	37.32	21.62	36.65	–	–
η	0.14	44.07	37.46	21.70	36.82	–	–
η	0.20	43.95	37.38	21.10	36.60	–	–
η	0.28	43.50	36.94	20.94	36.22	–	–
γ	0.00	43.55	37.12	20.10	36.08	–	–
γ	0.10	43.85	37.30	21.21	36.55	–	–
γ	0.18	44.07	37.46	21.70	36.82	–	–
γ	0.26	43.99	37.33	21.34	36.66	–	–
γ	0.36	43.55	36.86	20.88	36.21	–	–
R	4	43.34	36.91	20.82	36.10	2.7	-1.84
R	8	43.69	37.20	21.35	36.48	4.5	-1.71
R	12	43.92	37.38	21.58	36.70	5.8	-1.65
R	16	44.07	37.46	21.70	36.82	6.6	-1.63
R	24	44.05	37.43	21.68	36.80	6.9	-1.63

F.5 Efficiency and Complexity

The efficiency comparison isolates the recovery module after shared feature extraction. This measurement matches the main-paper claim: the method changes the semantic field before final prompt matching, rather than proposing a new feature extractor or a new language encoder.

Table A7: Relative accuracy–efficiency trade-off of the recovery modules. Measurements are taken after shared feature extraction under the same point count, batch size, prompt setting, and hardware. Absolute latency, peak memory, and throughput are reported separately in Table A8. Mean is mIoU (%), and Gain is relative to direct point–text matching.

Variant	Added params	Rel. latency	Rel. memory	Mean mIoU	Gain
Direct point–text matching	0.00M	1.00x	1.00x	32.72	–
Generic MLP refinement	0.42M	1.05x	1.04x	33.07	+0.35
Density-guided denoising only	0.58M	1.12x	1.09x	34.69	+1.97
Full model	**1.18M**	**1.31x**	**1.24x**	**36.82**	**+4.10**

Table A8: Measured absolute recovery-stage cost under the default evaluation configuration. Latency is measured in ms/sample, memory is peak allocated memory in MB, and throughput is measured in samples/s. The shared feature extractor and text encoder are excluded.

Variant	Latency (ms/sample)	Peak memory (MB)	Throughput (samples/s)
Direct point–text matching	4.82	412	207.5
Density-guided denoising only	5.91	468	169.2
Full model	**6.32**	**511**	**158.2**

Measurement boundary and interpretation. All measurements in this section exclude the shared feature extractor and text encoder. The direct baseline contains only masked point–text

matching, while the two reported recovery variants add the corresponding recovery computations before the same final matching rule. This separation makes the reported overhead attributable to semantic-field recovery rather than to a changed backbone or language interface. The Generic MLP row is retained in the relative comparison because its absolute timed-forward log is not part of the supplied measurement record; it is therefore omitted from the absolute-cost table rather than assigned an inferred latency.

F.5.1 INPUT-SIZE SCALING

To measure how computation changes with the number of input points, we keep the test samples, prompt set, valid-prompt mask, batch size, random seed, and evaluation code fixed and vary only the point count N. The preferred operating points are $N \in \{1024, 2048, 4096, 8192\}$; if the data pipeline supports fewer legal sizes, all available sizes are reported instead. Downsampling must use one fixed rule, and repeated padding is not used because it would obscure the actual scaling behavior.

For each N, we record mean latency, standard deviation, p95 latency, peak allocated memory, throughput, and mean mIoU. The same twenty warm-up iterations and at least one hundred timed forward passes are used for every point count. GPU synchronization is performed before and after each timed region, and peak-memory statistics are reset before each measurement. The hardware, CUDA/PyTorch versions, batch size, and point-count range are reported with the results.

Table A9: Measured recovery-stage scaling with input point count. Mean and p95 latency are reported in ms/sample, peak memory in MB, throughput in samples/s, and mIoU in percent.

N	Variant	Mean	Std.	p95	Mem.	Throughput	mIoU
1024	Direct	2.46	0.04	2.53	286	406.5	32.58
1024	Denoising	3.04	0.05	3.13	327	328.9	34.48
1024	Full	3.28	0.06	3.38	361	304.9	36.54
2048	Direct	4.82	0.07	4.95	412	207.5	32.72
2048	Denoising	5.91	0.09	6.08	468	169.2	34.69
2048	Full	6.32	0.10	6.51	511	158.2	36.82
4096	Direct	9.34	0.13	9.59	684	107.1	32.75
4096	Denoising	11.46	0.17	11.79	798	87.3	34.72
4096	Full	12.24	0.19	12.61	879	81.7	36.84
8192	Direct	18.42	0.26	18.93	1226	54.3	32.69
8192	Denoising	22.58	0.31	23.18	1454	44.3	34.65
8192	Full	24.16	0.36	24.86	1605	41.4	36.79

The measured scaling summary fits $\mathrm{latency}(N) \propto N^{\beta}$ and $\mathrm{memory}(N) \propto N^{\zeta}$ on the four point-count settings. Specifically, we fit an ordinary least-squares line to $(\log N, \log \text{ latency})$ and $(\log N, \log \text{ memory})$ over $N \in \{1024, 2048, 4096, 8192\}$ using the displayed means. The resulting latency exponents are approximately 0.97 for direct matching, 0.96 for denoising only, and 0.96 for the full model; the corresponding memory exponents are approximately 0.70, 0.72, and 0.72. These exponents summarize the observed implementation behavior rather than replacing an implementation-level complexity derivation, and the displayed precision can cause small differences from fits using unrounded logs.

F.5.2 TRUNCATION-LEVEL SCALING

The truncation level R controls the computational upper bound of the variational Dirichlet process. We reuse the existing $R \in \{4, 8, 12, 16, 24\}$ sensitivity sweep and report latency, memory, and throughput from the same measurement protocol. The default $R = 16$ provides the best measured mean mIoU while keeping the active-component count and runtime below the larger truncation setting.

Complexity accounting. Let N denote the number of input points, K the number of valid candidate prompts, k the local-neighbor count, d the point-feature or prompt-response dimension used by the denoising affinities, h the semantic-state dimension, R the truncation level, T the number of discrete Wiener correction steps, and L the number of recovery blocks. Direct masked point–text

Table A10: Measured truncation-level cost. The table couples the computational cost of R with active semantic components, normalized ELBO, and mean mIoU. Relative latency and memory are normalized to the default $R = 16$ setting.

R	Latency (ms)	Rel. latency	Peak memory (MB)	Rel. memory	Throughput	Active comps.	ELBO / Overall mIoU
4	5.62	0.89x	458	0.90x	177.9	2.7	-1.84 / 36.10
8	5.91	0.94x	474	0.93x	169.2	4.5	-1.71 / 36.48
12	6.14	0.97x	489	0.96x	162.9	5.8	-1.65 / 36.70
16	6.32	1.00x	511	1.00x	158.2	6.6	-1.63 / 36.82
24	6.83	1.08x	542	1.06x	146.4	6.9	-1.63 / 36.80

matching costs $O(NKd)$; with the feature dimension fixed, this is commonly written as $O(NK)$. Constructing prompt-conditioned local affinities and aggregating k neighbors costs $O(Nk(d+K))$. When dense $h \times h$ semantic covariances are formed and inverted for boundary projection, the corresponding additional upper bound is $O(Nkh^2 + Nh^3)$. Evaluating the variational score over R components for T correction steps costs $O(NTRh^2)$. Thus, a dense implementation of L recovery blocks has the following time-complexity upper bound:

$$\mathcal{O}\big(L\left[NKd + Nk(d+K) + Nkh^2 + Nh^3 + NTRh^2\right]\big). \tag{36}$$

Here the Nh^3 term is specific to dense per-point covariance inversion; diagonal or low-rank covariance implementations reduce this term. The peak inference storage, when blocks are processed sequentially, is

$$\mathcal{O}\big(N(d + K + k + h^2 + Th) + Rh^2\big), \tag{37}$$

where the first term stores point states, masked prompt responses, local neighborhoods, covariance/projection quantities, and the current correction path, while Rh^2 stores the truncated component statistics. Training that retains all L correction paths adds $O(LNTh)$ auxiliary storage; sequential inference does not multiply peak memory by L. Under the fixed d, h, k, R, T, and L used in the reported experiments, the time bound is approximately linear in N, consistent with the empirical exponents in Table A9. The latency and memory measurements exclude the shared feature extractor and text encoder and therefore quantify recovery-stage overhead only.

G Qualitative Visualizations

The following qualitative analyses complement the quantitative results in the main paper. The point-cloud visualization compares final predictions across methods, the t-SNE visualization tracks semantic-field states across recovery stages, the boundary-missing visualization complements the three deletion rates in Table A2, and the failure cases expose residual errors on challenging parts. External-baseline predictions in the boundary-missing visualization provide qualitative context; the formal quantitative comparison remains the aligned CoSMo3D–DGV-SFR evaluation.

The point-cloud examples are shown immediately below to compare queried-part localization across representative objects.

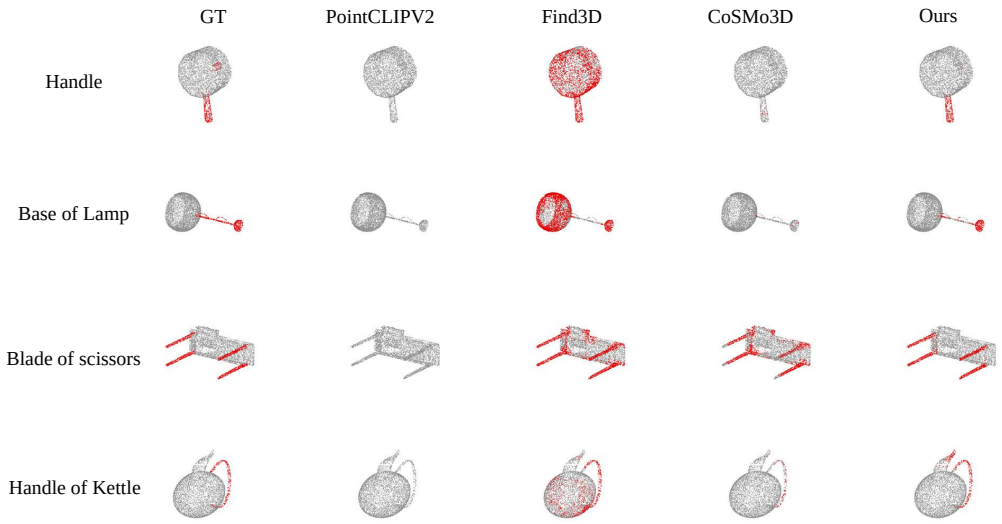

Figure A1: Qualitative point-cloud comparison on four part queries. Each row shows ground truth and predictions from PointCLIPV2, Find3D, CoSMo3D, and DGV-SFR. The examples provide a visual comparison of queried-part localization for a handle, a lamp base, a scissor blade, and a kettle handle.

The semantic-field states across recovery stages are shown next.

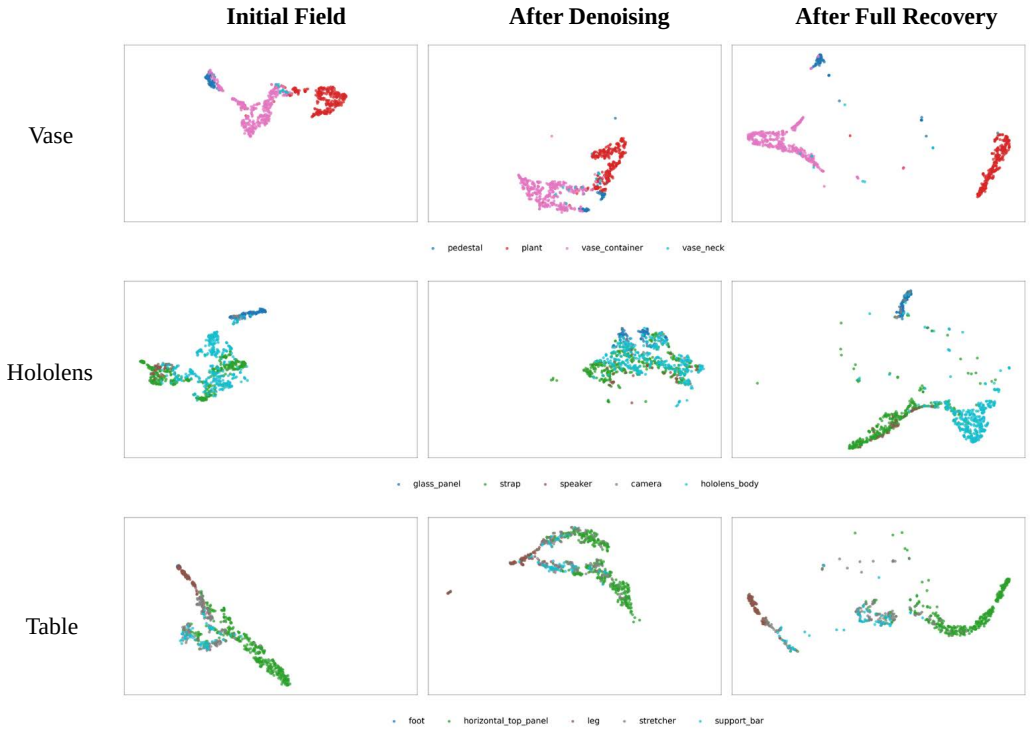

Figure A2: t-SNE visualization of the semantic field at three stages: the initial field, after density-guided denoising, and after full Dirichlet–Wiener recovery. Rows show Vase, HoloLens, and Table; colors identify the part labels listed in the corresponding row legend.

The boundary-missing visualization follows the stress-test description and displays the 8%, 12%, and 16% settings together.

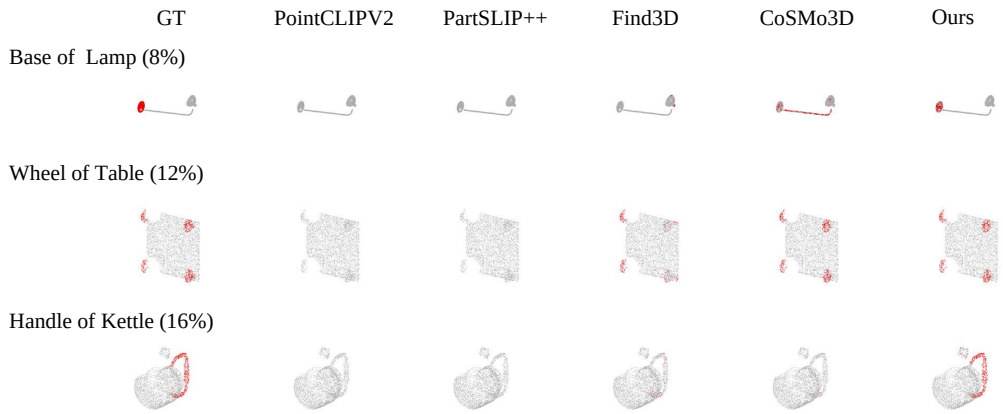

Figure A3: Qualitative comparison under boundary-missing inputs at 8%, 12%, and 16% removal. Rows show Base of Lamp, Wheel of Table, and Handle of Kettle, respectively; columns show ground truth and predictions from PointCLIPV2, PartSLIP++, Find3D, CoSMo3D, and DGV-SFR. This visualization complements the main-paper 8% quantitative result and the measured 12%/16% results in Table A2.

The remaining examples show representative failure cases.

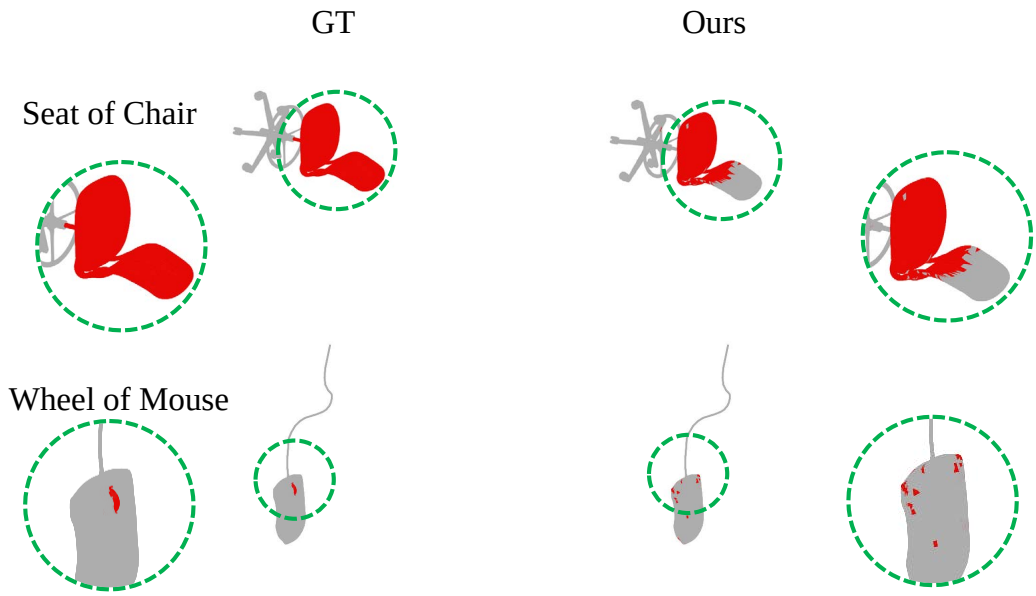

Figure A4: Representative challenging cases comparing ground truth with DGV-SFR predictions for Seat of Chair and Wheel of Mouse. Green dashed circles identify the queried regions for inspection; differences between the predicted and reference part extents illustrate residual errors on small or weakly separated structures.

H LIMITATIONS AND FAILURE MODES

The recovery framework is conservative by design, but it still depends on the information available in the initial prompt-conditioned semantic field. If a large connected region is assigned to an incorrect semantic mode with high confidence, the anchor term can preserve that state and reduce the corrective effect. If two adjacent parts are thin, geometrically weak, or described by ambiguous prompts, the boundary projection may not fully separate their semantic updates. Conversely, when the initial field is already clean, the recovery stages are expected to provide smaller gains because the reliability gates and anchor tethering limit unnecessary movement.

The method also introduces computation after shared feature extraction. The reported overhead is moderate for the evaluated benchmarks, but larger point clouds may require sparse neighborhoods, fewer recovery blocks, or low-rank approximations of P_i^ℓ. These limitations are consistent with the main-paper conclusion and motivate stronger boundary evidence, prompt disambiguation, and scalable projection approximations in future work.