# FACTORIZED PROBABILISTIC INFERENCE FOR MULTIMODAL FEW-SHOT POINT CLOUD SEGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal Few-Shot Point Cloud Semantic Segmentation (MFS-PCS) leverages Vision-Language Models (VLMs) to reduce reliance on dense 3D annotations. However, existing methods often perform inference in an entangled latent space that jointly mixes visual, textual, and geometric cues. Such a monolithic treatment obscures the distinct roles of semantic priors and geometric structure, often leading to biased prototypes and misalignment between external semantics and the intrinsic manifold of the query scene. From a probabilistic perspective, we argue that few-shot inference can be naturally factorized into semantic likelihood estimation and geometric manifold regularization. Motivated by this view, we propose Decoupling Semantic Likelihood and Geometric Priors (DSG). For semantic transfer, we formulate class-conditional multi-prototype estimation through Maximum Likelihood Estimation and instantiate it with FPS-initialized single-step Lloyd/Hard-EM refinement of sparse support observations. For geometric regularization, we introduce unsupervised structural anchors that capture the latent Voronoi structure of the query manifold. These anchors act as transductive geometric priors, encouraging predictions to respect local neighborhood structure and remain coherent across geometrically related regions. By separating what a category is from where and how it appears in the query geometry, DSG better aligns external multimodal semantics with internal 3D topology. Extensive experiments on S3DIS and ScanNet show that DSG achieves new state-of-the-art performance for MFS-PCS.

## 1 INTRODUCTION

3D scene semantic understanding (Qi et al., 2017a;b; Thomas et al., 2019; Hu et al., 2020; Zhan et al., 2023; Wu et al., 2024) is a fundamental capability for real-world perception systems (Xiao et al., 2024), including autonomous driving, robotics, and embodied intelligence. Although fully supervised methods have achieved substantial progress, their success relies heavily on dense point-level annotations, which are costly and labor-intensive to obtain. This limitation has motivated Few-Shot Point Cloud Semantic Segmentation (FS-PCS) (Zhao et al., 2021b; Ning et al., 2023; He et al., 2023; Wang et al., 2025b;a), which aims to segment novel categories using only a few annotated support samples.

Early FS-PCS studies mainly rely on 3D geometric features and support prototypes. However, few-shot supervision provides limited category evidence and is often insufficient to capture the large intra-class variation of real 3D scenes. Recent studies therefore turn to Multimodal FS-PCS (MFS-PCS), which incorporates 2D images and pre-trained Vision-Language Models (VLMs) (Radford et al., 2021; Li et al., 2022) as external semantic priors. These multimodal cues enrich sparse 3D support sets with high-level class knowledge and improve semantic transfer to novel categories. Nevertheless, existing methods, such as MM-FSS (An et al., 2025b), typically project textual, visual, and geometric cues into a shared latent space and perform inference in this entangled representation. Such a monolithic treatment overlooks a key distinction: semantic priors describe what a category is, while query geometry constrains where and how labels should vary in a scene.

This coupling is particularly problematic under few-shot supervision. Text embeddings provide abstract class-level semantics but lack spatial granularity, while support prototypes are estimated from very few samples and may not cover the query distribution (Liu et al., 2020). When query features are forced to align with these sparse and potentially biased external priors, the resulting decision



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.

boundaries can deviate from the intrinsic manifold of the query point cloud. As illustrated in Figure 1 (left), the conventional coupled paradigm mixes external semantics and internal geometry in a cluttered latent space, which can cause semantic misalignment and boundary distortion. We argue that this failure stems from treating two statistically different factors as one: semantic likelihood, which measures compatibility with multimodal category evidence, and geometric prior, which encodes the local structure of the query manifold.

From a probabilistic perspective, few-shot segmentation should infer the most probable label by jointly considering these two factors. The semantic likelihood evaluates how likely a query point belongs to a class given support, visual, and textual evidence. The geometric prior regularizes this prediction according to the topology of the query scene, encouraging nearby or structurally related points to share coherent labels. This view naturally suggests a factorized inference principle: semantic transfer and geometric regularization should be modeled separately and then combined for posterior prediction. As shown in Figure 1 (right), such factorization allows external multimodal semantics to guide category recognition without overriding the intrinsic geometric organization of the query set.

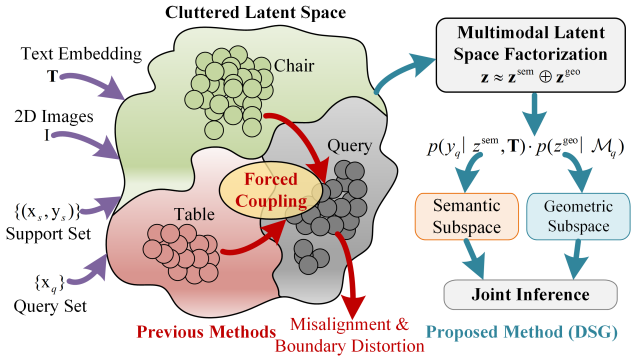



Figure 1: Conceptual overview. Existing MFS-PCS methods perform inference in a shared latent space that entangles multimodal semantics with query geometry. Such forced coupling distorts the intrinsic manifold structure and causes semantic misalignment under sparse supervision. In contrast, DSG factorizes inference into semantic likelihood estimation and geometric manifold regularization, enabling robust semantic transfer while preserving local geometric consistency.

Motivated by this principle, we propose Decoupling Semantic Likelihood and Geometric Priors (DSG), a probabilistic framework for MFS-PCS. DSG decouples latent reasoning through a dual-stream interaction network. First, for semantic likelihood estimation, we introduce the Tri-Modal Semantic Alignment (TMSA) stream. We formulate class-conditional multi-prototype estimation as a Maximum Likelihood Estimation (MLE) problem, implemented by FPS-initialized single-step Lloyd/Hard-EM refinement. The contribution lies in semantic–geometric inference rather than a new clustering update rule. These prototypes are further integrated with visual prototypes and text embeddings to estimate the semantic likelihood (Snell et al., 2017) of query points. Second, for geometric prior modeling, we propose the Structure Regularized Inference (SRI) stream based on unsupervised manifold regularization. It introduces structural anchors derived from competitive learning on the query set, which capture the latent topological structure of the query scene and provide a geometric basis for refining predictions. Finally, DSG performs maximum a posteriori inference by balancing semantic fidelity with geometric continuity, yielding predictions that are both semantically discriminative and geometrically coherent. Our main contributions are summarized as follows:

- We propose a probabilistic framework for MFS-PCS that factorizes few-shot inference into semantic likelihood estimation and geometric prior regularization, providing a principled formulation for decoupling semantic transfer from geometric reasoning.
- We reformulate support prototype construction as a Maximum Likelihood Estimation problem and introduce an EM-based prototype estimation strategy for robust semantic learning under sparse supervision. We further propose transductive manifold regularization with unsupervised structural anchors to preserve the intrinsic topology of the query scene.
- We develop the TMSA and SRI modules to implement the proposed factorized inference framework. Extensive experiments on S3DIS and ScanNet demonstrate state-of-the-art performance on MFS-PCS benchmarks.

## 2 Related Work

**Fully Supervised Point Cloud Semantic Segmentation.** PointNet (Qi et al., 2017a) pioneered point cloud semantic segmentation by introducing permutation-invariant operations for unordered point sets. Subsequent methods (Qi et al., 2017b; Thomas et al., 2019; Wang et al., 2019; Guo

et al., 2021; Kolodiazhnyi et al., 2024) improved segmentation performance through more effective local feature aggregation and geometric modeling. Many of these approaches rely on farthest point sampling with quadratic complexity, limiting their scalability to large-scale scenes. To address this issue, RandLA-Net (Hu et al., 2020) introduced efficient random sampling for large-scale point cloud processing, inspiring a series of lightweight segmentation frameworks. More recently, Transformer-based architectures (Zhao et al., 2021a; Lai et al., 2022; Wu et al., 2022; 2024) have demonstrated strong capability in capturing long-range geometric dependencies. Despite their success, fully supervised methods fundamentally depend on dense point-level annotations and lack the ability to generalize to novel categories with limited supervision.

**Few-Shot Point Cloud Semantic Segmentation.** Few-shot point cloud semantic segmentation (FS-PCS) aims to recognize novel categories from only a few annotated support samples, significantly reducing annotation costs while improving adaptability. The pioneering work attMPTI (Zhao et al., 2021b) introduced metric-based prototype learning for FS-PCS, and subsequent methods (Ning et al., 2023; He et al., 2023; Xu et al., 2023; Li et al., 2024; Wang et al., 2025b) mainly focus on reducing the distribution gap between support and query sets. However, the original attMPTI benchmark suffers from projection leakage and overly sparse point distributions, which limit its practicality in real-world scenarios. COSeg (An et al., 2024) addresses these issues by establishing a more realistic evaluation benchmark. Recent studies further extend FS-PCS to the multimodal setting. MFS-PCS (An et al., 2025b) incorporates 2D images and Vision-Language Models (VLMs) (An et al., 2025a) as external semantic priors to alleviate the semantic ambiguity caused by sparse 3D supervision. DA-FSS (Bian & Xu, 2026) decouples only at the network parameter level. While these methods demonstrate the effectiveness of multimodal semantic transfer, they typically rely heavily on semantic anchors in a shared latent space, often overlooking the intrinsic geometric structure of the query scene. Such entangled modeling can distort physical boundaries and weaken geometric consistency.

# 3 METHODOLOGY

## 3.1 PROBABILISTIC FORMULATION

**Few-shot Learning as Posterior Inference.** We frame Few-shot Point Cloud Semantic Segmentation (FS-PCS) within a probabilistic meta-learning framework (Grant et al., 2018; Finn et al., 2017). Let $\mathcal{D}$ be a distribution over tasks, where each task $\mathcal{T}_i \sim \mathcal{D}$ consists of a support set $\mathcal{S}_i$ and a query set $\mathcal{Q}_i$. The support set $\mathcal{S}_i = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{N\times K}$ contains point coordinates and features $\mathbf{x}_s \in \mathbb{R}^{M\times D_{in}}$ with corresponding dense labels $\mathbf{y}_s \in \{0, 1, \dots, N\}^M$. Our goal is to learn a parametric encoder $\mathcal{E}_\phi$ that maps inputs to a latent metric space $\mathcal{Z}$, such that the predictive posterior distribution of the query labels $p(\mathbf{y}_q|\mathbf{x}_q, \mathcal{S}_i; \phi)$ is maximized. Formally, we seek to optimize the expected log-likelihood over the task distribution:

$$\phi^* = \underset{\phi}{\operatorname{argmax}}\, \mathbb{E}_{\mathcal{T}_i\sim\mathcal{D}} \left[ \sum_{(\mathbf{x}_q,\mathbf{y}_q)\in\mathcal{Q}_i} \log p(\mathbf{y}_q|\mathbf{x}_q, \mathcal{S}_i; \phi) \right] \tag{1}$$

**Multimodal Latent Space Factorization.** In the Multimodal FS-PCS setting, we hypothesize that an optimal metric space $\mathcal{Z}$ must simultaneously satisfy semantic alignment constraints and geometric topological constraints. Rather than treating modalities in isolation, we first construct a unified latent representation $\mathbf{z} \in \mathcal{Z}$ by fusing the Intermodal Features (IF) and Unimodal Features (UF):

$$\mathbf{z} = \mathcal{F}_{\text{fusion}}(\mathcal{H}_{\text{IF}}(\mathbf{x}), \mathcal{H}_{\text{UF}}(\mathbf{x})) \tag{2}$$

To theoretically justify our subsequent refinement modules (TMSA and SRI), we posit that this fused representation $\mathbf{z}$ is not monolithic but can be factorized into two disentangled components (Bengio et al., 2013): $\mathbf{z} \approx \mathbf{z}^{\text{sem}} \oplus \mathbf{z}^{\text{geo}}$.

- **Semantic Component ($\mathbf{z}^{\text{sem}}$):** This component captures high-level category attributes. It is explicitly aligned with the pre-trained Vision-Language prior (Text Embeddings $\mathbf{T}$ (Radford et al., 2021)), maximizing the semantic likelihood $p(\mathbf{y}|\mathbf{z}^{\text{sem}}, \mathbf{T})$.
- **Geometric Component ($\mathbf{z}^{\text{geo}}$):** This component (Belkin et al., 2006) preserves the intrinsic manifold topology of the 3D data. It is constructed without query semantic labels and is regularized by the Structural Anchors $\mathbf{P}_q$ via unsupervised manifold learning, ensuring geometric consistency $p(\mathbf{z}^{\text{geo}}|\mathcal{M}_{\text{manifold}})$.

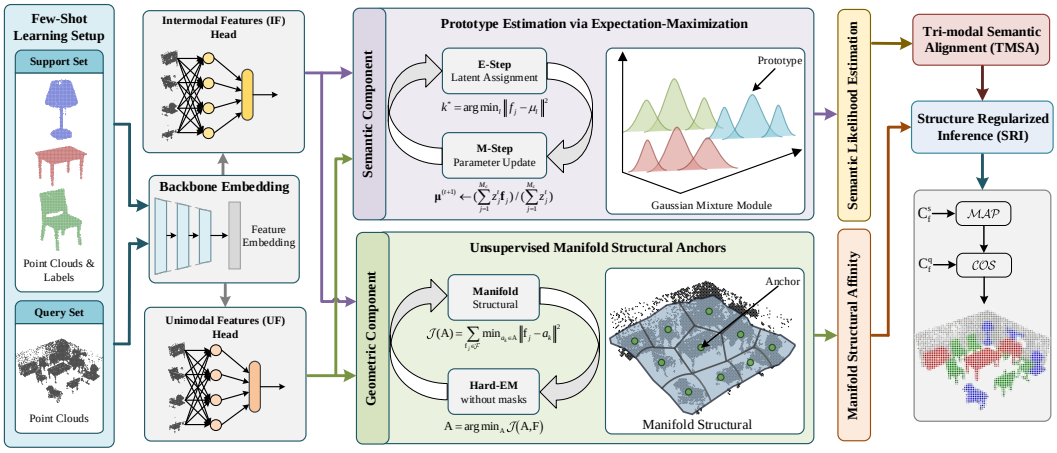



Figure 2: The DSG explicitly factorizes multimodal features into two decoupled inference streams. Top: the Semantic Flow (TMSA) performs EM-based prototype estimation, modeling class prototypes as Gaussian mixtures to maximize semantic likelihood under sparse supervision. Bottom: the Geometric Flow (SRI) employs unsupervised competitive learning to generate structural anchors that capture the latent Voronoi topology of the query manifold. Both prototype and anchor estimation use one FPS-initialized assignment-and-mean update; the cyclic arrows depict the update stages, not iteration to convergence. Final inference combines support-derived semantic evidence with query structural affinities.

Consequently, the final inference is modeled as a joint estimation over these factorized latent factors (Gordon et al., 2018):

$$p(\mathbf{y}_q|\mathbf{x}_q) \propto \underbrace{p(\mathbf{y}_q|\mathbf{z}^{\text{sem}}, \mathbf{T})}_{\text{Semantic Likelihood}} \cdot \underbrace{p(\mathbf{z}^{\text{geo}}|\mathcal{M}_q)}_{\text{Geometric Prior}} \tag{3}$$

We propose DSG, illustrated in Figure 2, a probabilistic framework that explicitly factorizes inference into two decoupled pathways. The probabilistic formulation describes the decomposition of semantic evidence and query structural regularization. Prototype and anchor estimation use classical Lloyd updates; their exact implementation is given in App. F.1.

### 3.2 Cross-modal Visual Fusion

Generally, a backbone network $\mathcal{E}_{backbone}$ is employed to extract high-dimensional latent features from raw point data, yielding $\mathbf{F}_s = \mathcal{E}_{backbone}(\mathbf{x}_s)$ and $\mathbf{F}_q = \mathcal{E}_{backbone}(\mathbf{x}_q)$. In a multimodal setting (An et al., 2025b), we utilize an IF Head $\mathcal{H}_{\text{IF}}$ and a UF Head $\mathcal{H}_{\text{UF}}$ to project these features into Intermodal Features (aligned with text/image) and Unimodal Features (point-specific), respectively:

$$\begin{aligned} \mathbf{F}_{\text{s}}^{\text{i}} &= \mathcal{H}_{\text{IF}}(\mathbf{F}_{\text{s}}), \quad \mathbf{F}_{\text{s}}^{\text{u}} = \mathcal{H}_{\text{UF}}(\mathbf{F}_{\text{s}}) \\ \mathbf{F}_{\text{q}}^{\text{i}} &= \mathcal{H}_{\text{IF}}(\mathbf{F}_{\text{q}}), \quad \mathbf{F}_{\text{q}}^{\text{u}} = \mathcal{H}_{\text{UF}}(\mathbf{F}_{\text{q}}) \end{aligned} \tag{4}$$

where $\mathbf{F}_{\text{s}}^{\text{i}} \in \mathbb{R}^{M_s \times D_{\text{t}}}$ and $\mathbf{F}_{\text{q}}^{\text{i}} \in \mathbb{R}^{M_q \times D_{\text{t}}}$ are the same dimension $D_t$ as the text embedding $\mathcal{T}$. Meanwhile, $\mathbf{F}_{\text{s}}^{\text{u}} \in \mathbb{R}^{M_s \times D}$ and $\mathbf{F}_{\text{q}}^{\text{u}} \in \mathbb{R}^{M_q \times D}$ retain the point-specific dimension $D$.

IF and UF are complementary representation heads, not separate semantic and geometric reasoning branches. Both representations are used downstream: TMSA performs support–query semantic matching, while SRI models query structural consistency.

**Prototype Estimation via Expectation-Maximization.** We formulate class-conditional multi-prototype estimation as a Maximum Likelihood Estimation (MLE) problem. We assume the support features for a specific category $c$ are drawn from a Mixture of Gaussians (MoG) distribution (Snell et al., 2017), where the latent semantic modes correspond to the prototypes $\mathbf{P}^c = \{\boldsymbol{\mu}_k^c\}_{k=1}^{N_p}$, representing the central modes of the category's feature distribution.

Formally, given the filtered support features $\mathbf{F}_{\text{s}}^{\text{i},c}$ and $\mathbf{F}_{\text{s}}^{\text{u},c}$ (denoted generally as $\mathcal{F}^c = \{\mathbf{f}_j\}_{j=1}^{M_c}$) extracted via the category mask $\mathbf{R}^c$, we aim to maximize the log-likelihood function:

$$\mathcal{L}(\mathbf{P}^c) = \sum_{j=1}^{M_c} \log \sum_{k=1}^{N_p} \pi_k \mathcal{N}(\mathbf{f}_j|\boldsymbol{\mu}_k^c, \boldsymbol{\Sigma}_k^c) \tag{5}$$

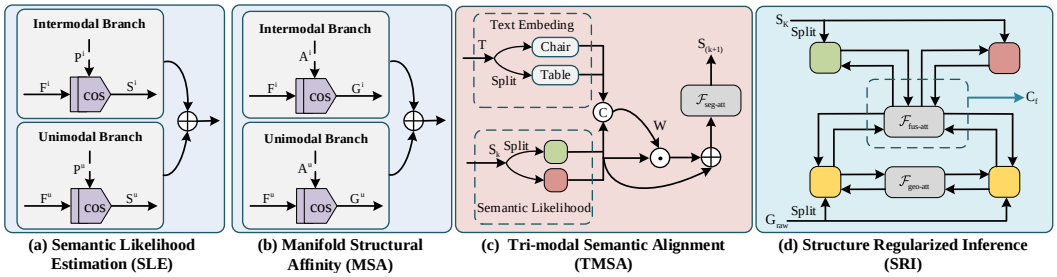



Figure 3: Detailed module structure. (a) SLE calculates the initial semantic probability by aggregating the cosine similarity between the IF/UF branches and the semantic anchors. (b) Similarly, MSA calculates geometric affinity using the learned structural anchors to capture popular topological structures. (c) TMSA iteratively refines semantic likelihood. It integrates text embeddings with visual feature semantic priors through the interactive attention mechanism $\mathcal{F}_{\mathrm{seg-att}}$. (d) SRI employs a cross-attention mechanism, utilizing the original geometric structure $\mathbf{G}_{\mathrm{raw}}$ to correct the semantic prediction $\mathbf{S}_{\mathrm{K}}$, ensuring that the prediction $\mathbf{C}_{\mathrm{f}}$ follows the intrinsic manifold of the query set.

where $\pi_k$ is the mixing coefficient. Direct optimization of $\mathcal{L}$ is intractable. We instantiate prototype refinement with a single Hard-EM assignment-and-mean update. The proof of Eq. (5) is in App. A.1.

*Initialization.* FPS is performed on the 3D support coordinates, and the selected feature vectors initialize the centroids $\boldsymbol{\mu}^{(0)}$. Each support shot contributes $N_p/K$ prototypes per class, concatenated to give $N_p = 50$ prototypes by default (App. F.1).

*Optimization.* With equal mixture weights and shared isotropic covariance $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, the limit $\sigma^2 \to 0$ gives hard assignments. We execute the following two steps once, numerically equivalent to one Lloyd update with the same initialization and features:

(i) **E-Step (Latent Assignment):** We estimate the posterior distribution of the latent variables. Under the hard-assignment assumption, this reduces to assigning each feature $\mathbf{f}_j$ to the nearest prototype in the metric space, minimizing the quantization error:

$$z_{j,k}^{(t)} = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_l \|\mathbf{f}_j - \boldsymbol{\mu}_l^{(t)}\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

(ii) **M-Step (Parameter Update):** We maximize the expected log-likelihood with respect to $\mathbf{P}^c$. This yields a closed-form update rule where the new prototype is the mean of the features assigned to it:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{j=1}^{M_c} z_{j,k}^{(t)} \mathbf{f}_j}{\sum_{j=1}^{M_c} z_{j,k}^{(t)}} \tag{7}$$

The proof of Eq. (6) and Eq. (7) are in App. A.1.1.

**Semantic Likelihood Estimation.** Based on the refined prototypes $\mathbf{P}^c$, we compute the semantic likelihood of query features. We formulate this as a metric-based similarity in the latent space. For both Intermodal (IF) and Unimodal (UF) branches, we calculate the cosine affinity matrices, which measure the similarity between query features and prototypes in the respective branches:

$$\mathbf{S}^{\mathrm{i}} = \mathrm{sim}(\mathbf{F}_{\mathrm{q}}^{\mathrm{i}}, \mathbf{P}^{\mathrm{i}}), \quad \mathbf{S}^{\mathrm{u}} = \mathrm{sim}(\mathbf{F}_{\mathrm{q}}^{\mathrm{u}}, \mathbf{P}^{\mathrm{u}}) \tag{8}$$

where $\mathrm{sim}(\cdot)$ denotes cosine similarity. $\mathbf{S}^{\mathrm{i}}, \mathbf{S}^{\mathrm{u}} \in \mathbb{R}^{M_q \times N_p}$ serve as the *discriminative semantic priors* derived solely from support supervision.

Parallel to the semantic flow, the MSA unit, as shown in Figure 3 (b), utilizes unsupervised structural anchors to fuse visual modal information. We also propose a Structure Regularized Inference (SRI) module, as shown in Figure 3 (d), which can act as a geometric corrector. SRI receives the refined semantic likelihood ($\mathbf{S}_{\mathrm{k+1}}$) and the original geometric affinity ($\mathbf{G}_{\mathrm{raw}} = \mathbf{G}^{\mathrm{i}} + \mathbf{G}^{\mathrm{u}}$), the sum of IF and UF point-to-anchor similarities. By employing a two-stream cross-attention mechanism ($\mathcal{F}_{\mathrm{fus-att}}$), semantic predictions are propagated along the geometric manifold, maintaining consistency in point predictions within the same structural unit.

**Unsupervised Manifold Regularization via Structural Anchors.** Standard few-shot methods rely solely on the semantic transfer $\mathbf{S}$. However, due to potential domain shifts, the empirical support distribution may not strictly align with the query distribution. To address this, we introduce a *Transductive Calibration* mechanism using Structural Anchors.

We assume the query features lie on a low-dimensional manifold $\mathcal{M}_{\mathrm{q}}$. We seek a set of structural anchors $\mathbf{A} = \{\mathbf{a}_k\}_{k=1}^{N_a}$ that minimizes the reconstruction error of the query manifold:

$$\mathcal{J}(\mathbf{A}) = \sum_{\mathbf{f}_j \in \mathcal{F}_{\mathrm{q}}} \min_{\mathbf{a}_k \in \mathbf{A}} \|\mathbf{f}_j - \mathbf{a}_k\|^2 \tag{9}$$

The proof of Eq. (9) is in App. A.2.

Minimizing this objective partitions the query manifold into Voronoi cells, where each anchor serves as a local geometric pivot. This is the classical squared-error vector-quantization objective. In implementation, anchors are refined by one Lloyd assignment-and-mean update from 3D FPS initialization:

$$\mathbf{A}^b = \mathrm{Lloyd}_1(\mathbf{F}_{\mathrm{q}}^b; \mathbf{A}^{b,(0)}), \quad b \in \{\mathrm{i}, \mathrm{u}\} \tag{10}$$

Here $\mathrm{Lloyd}_1$ denotes one assignment-and-mean update, and $\mathbf{A}^{b,(0)}$ contains features at the FPS-selected query coordinates. No query labels are used, and cluster indices are not segmentation outputs (App. F.1).

**Manifold Structural Affinity.** We rely on the *Manifold Smoothness Assumption*: points close in the geometric manifold should share similar representations. We quantify this by computing the affinity between query features and the Structural Anchors:

$$\mathbf{G}^{\mathrm{i}} = \mathrm{sim}(\mathbf{F}_{\mathrm{q}}^{\mathrm{i}}, \mathbf{A}^{\mathrm{i}}), \quad \mathbf{G}^{\mathrm{u}} = \mathrm{sim}(\mathbf{F}_{\mathrm{q}}^{\mathrm{u}}, \mathbf{A}^{\mathrm{u}}) \tag{11}$$

Here, $\mathbf{G}^{\mathrm{i}}, \mathbf{G}^{\mathrm{u}} \in \mathbb{R}^{M \times N_{\mathrm{a}}}$ serve as *generic structural priors*. This structural prior acts as a regularizer, ensuring prediction continuity. The total affinities are aggregated as $\mathbf{S}_{\text{raw}} = \mathbf{S}^{\mathrm{i}} + \mathbf{S}^{\mathrm{u}}$ and $\mathbf{G}_{\text{raw}} = \mathbf{G}^{\mathrm{i}} + \mathbf{G}^{\mathrm{u}}$.

### 3.3 Tri-modal Semantic Alignment

Specifically, in the SLE unit shown in Figure 3 (a), we first fuse the similarities $\mathbf{S}^{\mathrm{i}}$ and $\mathbf{S}^{\mathrm{u}}$ from different visual modalities. To further integrate multimodal knowledge, we designed the TMSA module shown in Figure 3 (c). This module takes the text embedding of the class name and the current visual likelihood map $\mathbf{S}_k$ as input. The likelihood $\mathbf{S}_{(k+1)}$ is refined through the multi-head attention block $\mathcal{F}_{\mathrm{seg-att}}$, making it more discriminative of inter-class ambiguity. To effectively fuse the modalities, we propose a text-guided attention mechanism. We view the category text embeddings $\mathcal{T}$ as a high-level *semantic prior*. First, we compute the correlation between the aligned features and the text priors:

$$\mathbf{T} = \mathbf{F}_{\mathrm{q}}^{\mathrm{i}} \cdot \mathcal{T}^{\top} \tag{12}$$

We then learn a point-wise adaptive gating function. Let $\mathbf{S}_0 = \mathbf{S}_{\text{raw}}$. The gate generates weights $\mathbf{W}_{\mathrm{q}}$ based on text affinity and the current semantic similarity $\mathbf{S}_{\mathrm{k}}$:

$$\mathbf{W}_{\mathrm{q}} = \mathcal{F}_{\mathrm{mlp}}(\mathrm{Concat}\,(\mathbf{T}, \mathbf{S}_{\mathrm{k}})) \tag{13}$$

followed by the linear attention refinement:

$$\begin{aligned} \mathbf{S}'_{\mathrm{k}} &= \mathbf{T} \odot \mathbf{W}_{\mathrm{q}} + \mathbf{S}_{\mathrm{k}} \\ \mathbf{S}_{\mathrm{k+1}} &= \mathbf{S}'_{\mathrm{k}} + \mathcal{F}_{\mathrm{mlp}}(\mathcal{F}_{\mathrm{seg-att}}(\mathbf{S}'_{\mathrm{k}})) \end{aligned} \tag{14}$$

This iterative process ($k = 0 \ldots K-1$) aligns the visual features with the semantic text space, yielding a robust affinity $\mathbf{S}_{\mathrm{K}}$.

### 3.4 Structure Regularized Inference

The final stage integrates the semantic affinity $\mathbf{S}_{\mathrm{K}}$ (Likelihood) with the geometric structural manifold $\mathbf{G} = \mathbf{G}_{\text{raw}}$ (Prior). First, we refine the structural prior $\mathbf{G}$ via a self-attention module to capture long-range geometric dependencies, yielding $\mathbf{G}' = \mathcal{F}_{\text{geo-att}}(\mathbf{G})$. We then concatenate the semantic and geometric relations:

$$\mathbf{C}_{\mathrm{f}} = \mathcal{F}_{\mathrm{fus-att}}(\mathbf{S}_{\mathrm{K}} \oplus \mathbf{G}') \tag{15}$$

This representation $\mathbf{C}_{\mathrm{f}}$ combines semantic likelihood and structural prior.

**Inference.** Analogous processing is applied to the support set to obtain $\mathbf{C}_{\mathrm{f}}^{\mathrm{s}}$. We compute the final rectified prototypes via Masked Average Pooling (MAP): $\hat{\mathbf{P}} = \mathcal{MAP}(\mathbf{C}_{\mathrm{f}}^{\mathrm{s}})$. Finally, we perform

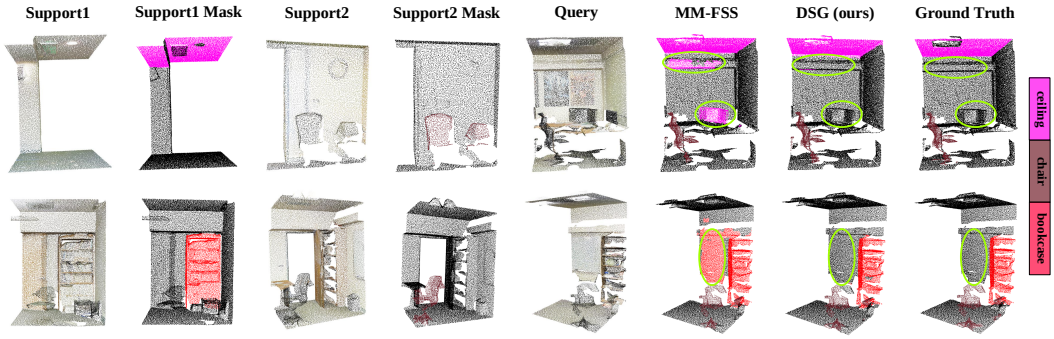



Figure 4: Qualitative comparison of 2-way 1-shot segmentation results on the S3DIS dataset. We compare DSG with the MM-FSS and the corresponding ground-truth. See circled marks for details.

Table 1: Comparison of few-shot quantitative results based on the S3DIS dataset. The best results are highlighted in bold. $S^0$ and $S^1$ denote two different splits. $^\dagger$ represents the introduction of multimodal data.

| **Methods** | **1-way 1-shot** | | | **1-way 5-shot** | | | **2-way 1-shot** | | | **2-way 5-shot** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S^0$ | $S^1$ | Mean | $S^0$ | $S^1$ | Mean | $S^0$ | $S^1$ | Mean | $S^0$ | $S^1$ | Mean |
| AttMPTI (Zhao et al., 2021b) | 36.32 | 38.36 | 37.34 | 46.71 | 42.70 | 44.71 | 31.09 | 29.62 | 30.36 | 39.53 | 32.62 | 36.08 |
| QGE (Ning et al., 2023) | 41.69 | 39.09 | 40.39 | 50.59 | 46.41 | 48.50 | 33.45 | 30.95 | 32.20 | 40.53 | 36.13 | 38.33 |
| QGPA (He et al., 2023) | 35.50 | 35.83 | 35.67 | 38.07 | 39.70 | 38.89 | 25.52 | 26.26 | 25.89 | 30.22 | 32.41 | 31.32 |
| COSeg (An et al., 2024) | 46.31 | 48.10 | 47.21 | 51.40 | 48.68 | 50.04 | 37.44 | 36.45 | 36.95 | 42.27 | 38.45 | 40.36 |
| COSeg$^\dagger$ (An et al., 2024) | 47.17 | 48.37 | 47.77 | 50.93 | 49.88 | 50.41 | 37.15 | 38.99 | 38.07 | 42.73 | 40.25 | 41.49 |
| MM-FSS (An et al., 2025b) | 49.84 | 54.33 | 52.09 | 51.95 | 56.46 | 54.21 | 41.98 | 46.61 | 44.30 | 46.02 | 54.29 | 50.16 |
| DA-FSS (Bian & Xu, 2026) | 49.14 | 55.93 | 52.54 | 51.90 | 60.45 | 56.18 | 42.07 | 47.16 | 44.62 | 45.78 | 56.30 | 51.04 |
| DSG (ours) | 52.48 | 59.78 | 56.13 | 54.45 | 65.24 | 59.85 | 50.22 | 54.30 | 52.26 | 48.99 | 61.27 | 55.13 |
| $\Delta$ vs. DA-FSS | +3.34 | +3.85 | +3.59 | +2.55 | +4.79 | +3.67 | +8.15 | +7.14 | +7.64 | +3.21 | +4.97 | +4.09 |

metric-based inference. The probability of a query point belonging to category $c$ is given by the softmax over the cosine distance to the rectified prototypes, scaled by a temperature $\tau$:

$$p(y = c|\mathbf{x}_q) = \frac{\exp(\tau \cdot \cos(\mathbf{C}_{\mathrm{f}}^{\mathrm{q}}, \hat{\mathbf{P}}_c))}{\sum_j \exp(\tau \cdot \cos(\mathbf{C}_{\mathrm{f}}^{\mathrm{q}}, \hat{\mathbf{P}}_j))} \tag{16}$$

This formulation can be interpreted to a maximum a posteriori estimation where the geometric structure acts as a regularizer, sharpening the decision boundaries in the latent space.

# 4 Experiments

## 4.1 Datasets and Evaluation

We evaluate on S3DIS and ScanNet using class-level splits with six and ten test classes per split, respectively; target test classes are unseen during training. We compute IoU for each target class, average over classes, and report the mean over test splits. App. F.2 gives class names, preprocessing, and episode construction. Main benchmark results quoted from prior papers are distinguished from the additional MM-FSS/DA-FSS checkpoint re-evaluations in App. F.3.

## 4.2 Implementation Details

The backbone network is the Stratified Transformer (Lai et al., 2022). Evaluation uses 100 episodes per class combination in the 2-way setting and 1000 episodes per class in the 1-way setting. Training samples 400 episodes per epoch. Scenes are partitioned into 1 m × 1 m blocks, with each block retaining at most 20,480 points (An et al., 2024). During visualization, the block size is 2 m × 2 m, and each block can retain up to 81,920 points. The pre-training phase is conducted for 100 epochs with a learning rate of 0.006, followed by a meta-learning phase of 30 epochs with a learning rate of 0.0001. The TACC (An et al., 2025b) inference is employed during testing. $\tau$ is set to 10 by default (He et al., 2023). All experiments were conducted on 4 × RTX A6000 GPUs.

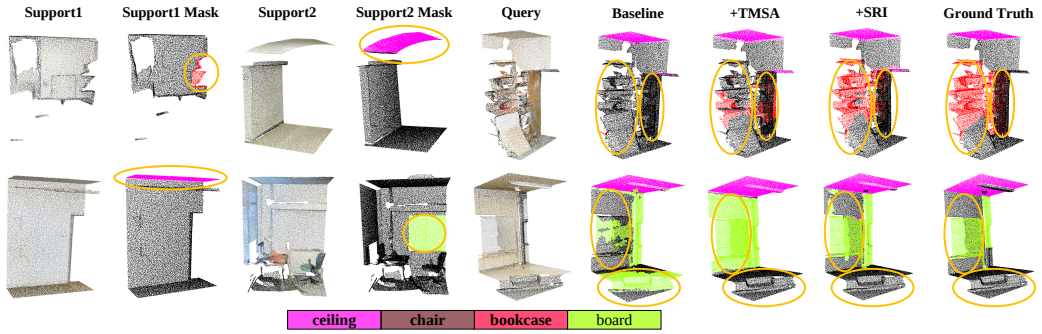



Figure 5: Qualitative ablation analysis on the S3DIS dataset. The progressive improvements brought by our proposed module are visualized, including the semantic likelihood introduced by adding TMSA and the geometric prior implemented by adding SRI.

Table 2: Comparison of few-shot quantitative results based on the ScanNet dataset. The best results are highlighted in bold.

| **Methods** | **1-way 1-shot** | | | **1-way 5-shot** | | | **2-way 1-shot** | | | **2-way 5-shot** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S^0$ | $S^1$ | Mean | $S^0$ | $S^1$ | Mean | $S^0$ | $S^1$ | Mean | $S^0$ | $S^1$ | Mean |
| AttMPTI (Zhao et al., 2021b) | 34.03 | 30.97 | 32.50 | 39.09 | 37.15 | 38.12 | 25.99 | 23.88 | 24.94 | 30.41 | 27.35 | 28.88 |
| QGE (Ning et al., 2023) | 37.38 | 33.02 | 35.20 | 45.08 | 41.89 | 43.49 | 26.85 | 25.17 | 26.01 | 28.35 | 31.49 | 29.92 |
| QGPA (He et al., 2023) | 34.57 | 33.37 | 33.97 | 41.22 | 38.65 | 39.94 | 21.86 | 21.47 | 21.67 | 30.67 | 27.69 | 29.18 |
| COSeg (An et al., 2024) | 41.73 | 41.82 | 41.78 | 48.31 | 44.11 | 46.21 | 28.72 | 28.83 | 28.78 | 35.97 | 33.39 | 34.68 |
| COSeg$^\dagger$ (An et al., 2024) | 41.95 | 42.07 | 42.01 | 48.54 | 44.68 | 46.61 | 29.54 | 28.51 | 29.03 | 36.87 | 34.15 | 35.51 |
| MM-FSS (An et al., 2025b) | 46.08 | 43.37 | 44.73 | 54.66 | 45.48 | 50.07 | 43.99 | 34.43 | 39.21 | 48.86 | 39.32 | 44.09 |
| DA-FSS (Bian & Xu, 2026) | 48.95 | 41.96 | 45.46 | 55.40 | 46.30 | 50.85 | 45.01 | 34.53 | 39.77 | 50.61 | 40.45 | 45.53 |
| DSG (ours) | 49.35 | 45.23 | 47.29 | 59.14 | 53.21 | 56.18 | 50.58 | 40.44 | 45.51 | 57.67 | 50.10 | 53.89 |
| $\Delta$ vs. DA-FSS | +0.40 | +3.27 | +1.83 | +3.74 | +6.91 | +5.33 | +5.57 | +5.91 | +5.74 | +7.06 | +9.65 | +8.36 |

### 4.3 Comparisons with State-Of-The-Arts

**S3DIS.** Table 1 presents the quantitative comparison results on the S3DIS (Armeni et al., 2016) dataset. DSG consistently outperforms all competing methods, establishing a new state-of-the-art (SOTA). DSG achieves substantial gains compared to the multimodal methods MM-FSS and DA-FSS. DSG increases mean mIoU by over 4% across all four settings by explicitly decomposing semantic and geometric reasoning. This significant difference validates our hypothesis that latent space decoupling is crucial for mitigating the misalignment between semantic priors and geometric structure.

**ScanNet.** To further evaluate the generalization ability of our framework across different scenarios, we report the performance on the ScanNet (Dai et al., 2017) dataset in Table 2. DSG maintains its advantage and continues to outperform state-of-the-art methods. In the 1-way setting, it outperforms the multimodal methods MM-FSS and DA-FSS by an average of over 4%. In the 2-way setting, it outperforms DA-FSS by an average of over 7%, and achieves a gain of over 8% in the 2-way 5-shot. This result demonstrates that the proposed DSG effectively captures the intrinsic manifold topology, even in noisy point clouds where purely semantic methods frequently fail.

**Visualization.** To visually validate the effectiveness of the decoupling strategy, as observed in Figure 4, the baseline method, constrained by the coupling latent space, often struggles with geometric ambiguity. For example, in the first row, MM-FSS cannot distinguish between the ceiling structure and the background; in the second row, MM-FSS cannot distinguish between the bookshelf structure and the background, resulting in a noisy segmentation image. Conversely, DSG generates high-fidelity predictions. Visualization results demonstrate that DSG's introduction of geometric priors significantly mitigates structural distortion and aligns predictions with the physical reality of the 3D scene. It's worth noting that DSG also achieved exciting results with AttMPTI's FS-PCS settings. See the App. C for details.

**Additional evaluations.** With fixed trained models, DSG has the highest mean mIoU over three independent episode sets in all eight settings (App. G.1). Adding SRI to MM-FSS and DA-FSS improves S3DIS $S^0$ 2-way 1-shot mIoU by 6.61 and 1.91 points after retraining (App. G.2). Boundary results for five neighborhood sizes and input-size/FPS timing are reported in Apps. G.3 and G.4.

### 4.4 Ablation Study

Unless otherwise specified, all ablation experiments were conducted under the 2-way 1-shot ($S^0$) setting on the S3DIS dataset. For more ablation studies, please refer to App. D.

**Impact of Multi-modal Components.** To assess the contribution of each modality in the TMSA, we conduct a component-wise analysis, as summarized in Table 3. Using only single-modal features (UF) from the point cloud backbone network yielded a baseline mIoU of 43.46%. Integrating text embeddings significantly improves performance to 46.62% (+3.16%), while adding inter-modal features (IF) from 2D images yields only a 0.73% gain. The complete DSG, which fuses UF, IF, and Text, achieves a peak mIoU of 50.22%. Adding IF to the text-aware UF model improves mIoU from 46.62 to 50.22 (+3.60 points), showing their complementary contribution. These are incremental modality ablations, not an assignment of IF exclusively to semantics or UF exclusively to geometry.

Table 3: Impact of Multi-modal Components.

| UF | IF | Text | mIoU |
|---|---|---|---|
| ✓ | | | 43.46 |
| ✓ | | ✓ | 46.62 |
| ✓ | ✓ | | 44.19 |
| ✓ | ✓ | ✓ | **50.22** |

Table 4: Effectiveness of Decoupled Streams.

| TMSA | SRI | mIoU |
|---|---|---|
| | | 38.63 |
| ✓ | | 42.90 |
| ✓ | ✓ | **50.22** |

**Effectiveness of Decoupled Streams.** To evaluate our decomposition hypothesis, as shown in Table 4. The baseline model achieves an mIoU of 38.63%. Incorporating TMSA yields a performance gain of 4.27%. The integration of SRI resulted in a significant performance leap of 7.32%, far exceeding the gain from semantic alignment. When combined, DSG achieves an overall improvement of 11.59% over the baseline. This finding confirms that maximizing semantic likelihood and enforcing geometric priors are complementary objectives, and that their joint inference is essential for achieving state-of-the-art performance.

To verify the specific contributions of the decomposed components, we present a visual ablation analysis in Figure 5. Compared with baseline, TMSA incorporates semantic likelihood, significantly reducing category ambiguity. For instance, in the first row, our method correctly identifies the salient bookcase region; in the second row, the baseline mistakenly segments the background into boards, which TMSA promptly corrects. These observations support our hypothesis that semantic alignment alone overlooks inherent topological constraints. To address this issue, SRI introduces geometric priors derived from unsupervised structural anchors, enabling semantic confidence to propagate along the underlying manifold and further refine the segmentation. As shown in the first-row bookcase and second-row board examples, SRI connects fragmented predictions into a coherent geometric surface.

**Additional analyses.** Low CKA/RSA scores describe representation similarity, not independence (App. D.1, Table 8). We use one TMSA layer (App. D.4, Table 11) and $N_p = N_a = 50$; full sensitivity results and the accuracy–efficiency analysis are provided in Apps. D.5 and G.4.

**Model Complexity Analysis.** As shown in Table 5, although the DSG increases the number of parameters, the parallelized design results in only a marginal increase in inference time (0.06 s), while GPU memory usage remains comparable to the baseline. These costs are acceptable considering the significant improvements.

Table 5: Model Complexity Analysis (RTX A6000).

| Methods | Params | GPU | Test |
|---|---|---|---|
| MM-FSS | **11.43M** | **23.56G** | **0.86s** |
| DSG (ours) | 16.01M | 24.73G | 0.92s |

## 5 Conclusion

This paper proposes a novel probabilistic framework, DSG, for MFS-PCS. Unlike traditional coupled feature learning paradigms, we explicitly decouple the inference process into semantic likelihood estimation and geometric prior modeling. We propose a Trimodal Semantic Alignment module that robustly estimates prototypes, and a Structure Regularized Inference module that captures manifold features of the query set. Extensive experiments on the S3DIS and ScanNet benchmark datasets demonstrate that DSG significantly outperforms state-of-the-art methods. Our work provides a perspective on how to effectively combine high-level semantic priors with low-level geometric structures in 3D understanding tasks. We hope this decomposition strategy will inspire future research in data-efficient 3D perception and multimodal representation learning.

### AI USE STATEMENT

Generative AI tools, including OpenAI Codex, were used to assist with language editing and manuscript revision. Codex also assisted with LaTeX template conversion, layout adjustments, and compilation. The authors take responsibility for the final text, claims, and reported results.

### ETHICS STATEMENT

This work studies multimodal few-shot point cloud segmentation using the existing S3DIS and ScanNet benchmarks, without collecting new data or conducting human-subject studies. The method may reduce annotation effort for 3D scene understanding. However, dataset bias and reliance on pretrained vision-language models may lead to uneven performance across scene types and categories. Sparse observations, ambiguous geometry, and domain shift can cause segmentation errors; deployment in robotics or other safety-sensitive applications requires application-specific validation and safeguards. Reuse of indoor-scene data should respect dataset access conditions and privacy protections. Limitations and failure cases are discussed in Appendices B.1 and G.3.

### REPRODUCIBILITY STATEMENT

The model formulation and inference procedure are described in Section 3, and the datasets, evaluation protocol, and principal training settings in Sections 4.1 and 4.2. Appendix A provides theoretical derivations and assumptions. Appendix F details the single-step Lloyd/Hard-EM implementation, preprocessing, class splits, episode sampling, baseline evaluation settings, and representation-similarity measures. Additional ablations and sensitivity analyses appear in Appendix D. Appendix G reports stability across independently sampled test episodes, SRI transfer experiments, boundary evaluation, and FPS and inference efficiency.

## REFERENCES

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3996–4006, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference*, 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. In *International Conference on Learning Representations*, 2025b.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1534–1543, 2016.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7 (11), 2006.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Wentao Bian and Fenglei Xu. Rethinking multimodal few-shot 3d point cloud segmentation: From fused refinement to decoupled arbitration. *arXiv preprint arXiv:2601.01456*, 2026.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations*, 2018.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

Shuting He, Xudong Jiang, Wei Jiang, and Henghui Ding. Prototype adaption and projection for few-and zero-shot 3d point cloud semantic segmentation. *IEEE Transactions on Image Processing*, 32:3199–3211, 2023.

Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11108–11117, 2020.

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20943–20953, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMlR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:249, 2008.

Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8500–8509, 2022.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Zhaoyang Li, Yuan Wang, Wangkai Li, Rui Sun, and Tianzhu Zhang. Localization and expansion: A decoupled framework for point cloud few-shot semantic segmentation. In *European Conference on Computer Vision*, pp. 18–34. Springer, 2024.

Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *European Conference on Computer Vision*, pp. 741–756. Springer, 2020.

Zhenhua Ning, Zhuotao Tian, Guangming Lu, and Wenjie Pei. Boosting few-shot 3d point cloud segmentation via query-guided enhancement. In *ACM International Conference on Multimedia*, pp. 1895–1904, 2023.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30, 2017b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *IEEE/CVF International Conference on Computer Vision*, pp. 6411–6420, 2019.

Changshuo Wang, Xiang Fang, and Prayag Tiwari. Dypolyseg: Taylor series-inspired dynamic polynomial fitting network for few-shot point cloud semantic segmentation. In *International Conference on Machine Learning*, 2025a.

Changshuo Wang, Shuting He, Xiang Fang, Meiqing Wu, Siew-Kei Lam, and Prayag Tiwari. Taylor series-inspired local structure fitting network for few-shot point cloud semantic segmentation. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 7527–7535, 2025b.

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 Transactions on Graphics*, 38(5):1–12, 2019.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4840–4851, 2024.

Aoran Xiao, Xiaoqin Zhang, Ling Shao, and Shijian Lu. A survey of label-efficient deep learning for 3d point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9139–9160, 2024.

Yating Xu, Conghui Hu, Na Zhao, and Gim Hee Lee. Generalized few-shot point cloud segmentation via geometric words. In *IEEE/CVF International Conference on Computer Vision*, pp. 21506–21515, 2023.

Lixin Zhan, Wei Li, and Weidong Min. Fa-resnet: Feature affine residual network for large-scale point cloud segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 118: 103259, 2023.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021a.

Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8873–8882, 2021b.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Jiaming Liu, Han Xiao, Chaoyou Fu, Hao Dong, and Peng Gao. No time to train: Empowering non-parametric networks for few-shot 3d scene segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3838–3847, 2024.

# A Theoretical Analysis and Proofs

In this appendix, we provide the rigorous theoretical derivations supporting the Probabilistic Formulation presented in Section 3.1. Specifically, we detail the derivation of the Expectation-Maximization algorithm for prototype estimation and derive its nearest-prototype assignment and mean update under the vanishing-variance assumption (Hard-EM). The implementation executes one such cycle, as detailed in App. F.1. Furthermore, we provide a geometric justification for the Structural Anchors.

## A.1 Derivation of Prototype Estimation via EM

**Problem Setup.** As defined in the methodology, we assume the support features $\mathcal{F}^c = \{\mathbf{f}_j\}_{j=1}^{M_c}$ for a specific category $c$ are generated from a Gaussian Mixture Model (GMM) with $N_p$ components (prototypes). The probability density function is given by:

$$p(\mathbf{f}_j|\Theta) = \sum_{k=1}^{N_p} \pi_k \mathcal{N}(\mathbf{f}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{17}$$

where $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{N_p}$ represents the model parameters. Our objective is to maximize the log-likelihood of the observed support data:

$$\mathcal{L}(\Theta) = \sum_{j=1}^{M_c} \log \left( \sum_{k=1}^{N_p} \pi_k \mathcal{N}(\mathbf{f}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{18}$$

**Lemma 1 (Evidence Lower Bound).** Direct maximization of $\mathcal{L}(\Theta)$ is intractable due to the summation inside the logarithm. We introduce latent variables $\mathbf{Z} = \{z_{j,k}\}$, where $z_{j,k} \in \{0, 1\}$ indicates whether feature $\mathbf{f}_j$ belongs to component $k$. By applying Jensen's Inequality, we derive the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\Theta) \geq \sum_{j=1}^{M_c} \sum_{k=1}^{N_p} \gamma_{j,k} \left( \log \pi_k + \log \mathcal{N}(\mathbf{f}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \triangleq Q(\Theta, \Theta^{(t)}) \tag{19}$$

where $\gamma_{j,k} = p(z_{j,k} = 1|\mathbf{f}_j, \Theta^{(t)})$ is the posterior probability (responsibility) computed in the previous iteration.

### A.1.1 Derivation of the Lloyd/Hard-EM Update Rule

We now prove that the update rules described in the main paper (Eq. 6 and Eq. 7) correspond to the analytical solution of maximizing this lower bound under the limit of isotropic, vanishing variance.

**Step 1: The E-Step (General Case).** In the standard EM algorithm, the responsibility $\gamma_{j,k}$ is computed using Bayes' theorem:

$$\gamma_{j,k} = \frac{\pi_k \exp\left(-\frac{1}{2}(\mathbf{f}_j - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{f}_j - \boldsymbol{\mu}_k)\right)}{\sum_{l=1}^{N_p} \pi_l \exp\left(-\frac{1}{2}(\mathbf{f}_j - \boldsymbol{\mu}_l)^\top \boldsymbol{\Sigma}_l^{-1}(\mathbf{f}_j - \boldsymbol{\mu}_l)\right)} \tag{20}$$

**Step 2: The Hard-EM Limit (Hard Assignment).** We simplify the model with two assumptions: (1) equal priors $\pi_k = 1/N_p$, and (2) isotropic covariance matrices with identical scalar variance $\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}$. The exponent term simplifies to $-\frac{1}{2\sigma^2}\|\mathbf{f}_j - \boldsymbol{\mu}_k\|^2$.

Consider the limit as $\sigma^2 \to 0$. Let $k^* = \operatorname{argmin}_l \|\mathbf{f}_j - \boldsymbol{\mu}_l\|^2$ be the index of the closest prototype. The behavior of $\gamma_{j,k}$ is:

$$\lim_{\sigma^2 \to 0} \gamma_{j,k} = \begin{cases} 1 & \text{if } k = k^* \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

*Proof.* For any $k \neq k^*$, the ratio of probabilities is dominated by the exponential term $\exp(-\frac{1}{2\sigma^2}(\|\mathbf{f}_j - \boldsymbol{\mu}_k\|^2 - \|\mathbf{f}_j - \boldsymbol{\mu}_{k^*}\|^2))$. Since the difference in squared distances is positive, this term tends to 0 exponentially fast as $\sigma^2 \to 0$. This rigorously justifies the discrete assignment indicator $z_{j,k}^{(t)}$ defined in Eq. 6 of our methodology.

**Step 3: The M-Step (Parameter Update).** To find the optimal prototypes $\boldsymbol{\mu}_k$, we maximize the Q-function with respect to $\boldsymbol{\mu}_k$ by setting the gradient to zero:

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_k} = \sum_{j=1}^{M_c} \gamma_{j,k} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left( -\frac{1}{2\sigma^2} \|\mathbf{f}_j - \boldsymbol{\mu}_k\|^2 \right) = 0 \tag{22}$$

Solving for $\boldsymbol{\mu}_k$:

$$\sum_{j=1}^{M_c} \gamma_{j,k} (\mathbf{f}_j - \boldsymbol{\mu}_k) = 0 \implies \boldsymbol{\mu}_k = \frac{\sum_{j=1}^{M_c} \gamma_{j,k} \mathbf{f}_j}{\sum_{j=1}^{M_c} \gamma_{j,k}} \tag{23}$$

Substituting the hard assignment limit $\gamma_{j,k} \to z_{j,k}^{(t)}$, we recover the exact update rule presented in Eq. 7:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{\mathbf{f}_j \in \mathcal{C}_k} \mathbf{f}_j}{|\mathcal{C}_k|} \tag{24}$$

where $\mathcal{C}_k$ denotes the set of features assigned to prototype $k$.

**Conclusion.** Under equal mixture weights, shared isotropic covariance, and the zero-variance limit, the assignment and mean update correspond to a Lloyd/Hard-EM refinement for a class-conditional GMM. The current implementation performs one update cycle rather than iterating to convergence.

### A.2 Justification of Unsupervised Manifold Regularization

Here we provide the theoretical justification for the objective function $\mathcal{J}(\mathbf{A})$ used in the Structural Anchors module (Eq. 9).

**Proposition 1 (Manifold Approximation via Quantization).** Let $\mathcal{M}$ be a smooth Riemannian manifold embedded in $\mathbb{R}^D$. The problem of finding a discrete set of anchors $\mathbf{A} = \{\mathbf{a}_k\}_{k=1}^{N_a}$ that best preserves the geometry of $\mathcal{M}$ is equivalent to minimizing the *Distortion Error* of the quantization mapping $q : \mathcal{M} \to \mathbf{A}$.

*Proof.* The expected distortion error for representing the manifold is defined as:

$$\mathcal{D}(\mathbf{A}) = \int_{\mathcal{M}} \|\mathbf{x} - q(\mathbf{x})\|^2 dP(\mathbf{x}) \tag{25}$$

where $q(\mathbf{x}) = \operatorname{argmin}_{\mathbf{a} \in \mathbf{A}} \|\mathbf{x} - \mathbf{a}\|$ maps a point to its nearest anchor. Approximating the integral with the empirical distribution of the query set $\mathcal{F}_q = \{\mathbf{f}_j\}_{j=1}^{M_q}$, we obtain the discrete objective:

$$\hat{\mathcal{D}}(\mathbf{A}) \approx \frac{1}{M_q} \sum_{j=1}^{M_q} \min_k \|\mathbf{f}_j - \mathbf{a}_k\|^2 \tag{26}$$

This is exactly proportional to the objective function $\mathcal{J}(\mathbf{A})$ defined in our method. The implementation uses one Lloyd assignment-and-mean update to refine the FPS-initialized anchors for this quantization objective (App. F.1). This implies that the affine graph constructed via $\mathbf{G} = \operatorname{sim}(\mathbf{F}_q, \mathbf{A})$ approximates the discrete Laplace-Beltrami operator on the manifold, thereby enforcing valid smoothness constraints on the semantic predictions and mitigating the effect of outliers during feature fusion.

## B Discussion

The effectiveness of probabilistic decomposition. The core premise of this work is that the "entanglement" of semantic and geometric features in a single latent space is the main bottleneck for few-shot point cloud segmentation. Our experimental results on the S3DIS and ScanNet datasets strongly support this hypothesis. By formally decomposing the posterior probability into semantic likelihood and geometric prior, DSG effectively decouples the challenges of "what to segment" (classification) and "where to segment" (boundary delineation). The quantitative advantages of our approach demonstrate that imposing explicit structural constraints is a more efficient data strategy

than indiscriminately fusing multimodal features, which often leads to overfitting in situations where data is scarce.

The synergistic effect between likelihood and prior. Our ablation experiments reveal a compelling interaction between the two proposed models. The TMSA module, based on EM prototype estimation, acts as a global semantic anchor, robustly identifying target object regions and suppressing background interference (as shown in the qualitative results). However, semantic features are inherently coarse. The SRI module compensates for this deficiency by acting as a local geometric corrector. It leverages unsupervised structural anchors to enforce smoothing constraints on the inherent manifold of the query point cloud. This synergy ensures that our predictions are not only semantically consistent with the supporting samples, but also geometrically faithful to the underlying 3D topology.

### B.1 LIMITATIONS

Although our framework has reached a new level of advancement, some limitations still exist. First, relying on pre-trained visual language models (such as CLIP) for text and image embedding makes us dependent on the quality of these external priors. In scenarios where the pre-trained data differs from the 3D point cloud domain (e.g., medical imaging or specialized industrial parts), semantic alignment accuracy may degrade. Secondly, the FPS initialization, single assignment-and-mean update, and affinity computations introduce additional overhead compared with simple prototype averaging. Finally, our geometric prior relies on the assumption that semantic boundaries align with geometric discontinuities (captured by Voronoi units). While this assumption generally holds, it is challenging when semantic variations occur on smooth surfaces (e.g., patterns on a planar wall). Specifically as follows:

**Spatial Context Ambiguity:** The model still has limitations in distinguishing adjacent objects with similar local geometry. For example, in rows 2 and 6 of Figure 6, DSG easily missegments a beam close to the ceiling as a ceiling.

**Sensitivity to Scale and Sparsity:** When the geometric space occupied by foreground objects is extremely small, they are easily ignored by the model (the missing board class in row 6). This exposes the limitation of the current mechanism: when the local point cloud is extremely sparse, structural anchor points degenerate into isolated points, failing to effectively extract meaningful topological information.

**Boundary Misalignment:** In rows 4 and 5, DSG can accurately identify the main structure of the bookshelf, but the segmentation at the edges is still not smooth enough. When semantic boundaries are inconsistent with geometric discontinuities (e.g., smooth surface boundaries based on texture or color), the manifold smoothness assumption underlying the SRI module will fail to some extent.

**Implications for Future Work:** While the SRI module significantly alleviates the large-area foreground leakage problem in previous methods (such as MM-FSS), we plan to explore density-adaptive anchors for the aforementioned extreme sparsity and boundary failure scenarios. Furthermore, exploring computationally efficient full-scene transductive segmentation (to replace the current block-based processing paradigm) is also a highly challenging and valuable future direction.

## C MORE QUANTITATIVE EXPERIMENTS

Table 6: Comparison of few-shot quantitative results based on the AttMPTI setting (2,048 Points). The best results are highlighted in bold. ($S^0$)

| **Methods** | **2-way 1-shot** | | **2-way 5-shot** | |
|---|---|---|---|---|
| | S3DIS | ScanNet | S3DIS | ScanNet |
| AttMPTI (Zhao et al., 2021b) (CVPR 2021) | 53.77 | 42.55 | 61.67 | 54.00 |
| Seg-PN (Zhu et al., 2024) (CVPR 2024) | 64.84 | 63.15 | 67.63 | 67.08 |
| TaylorSeg-PN (Wang et al., 2025b) (AAAI 2025) | 67.12 | 67.52 | 70.44 | 68.39 |
| DSG (ours) | **69.61** | **69.48** | **71.05** | **72.10** |

To eliminate the gain artifacts associated with "massive number of points" and "heavyweight networks," DSG switched to the AttMPTI setting (2,048 points and DGCNN). DSG still achieved a significant leap under this configuration, significantly surpassing the latest TaylorSeg-PN (Table 6). This fully demonstrates that DSG does not rely on massive numbers of points, its decoupling mechanism is independent of the backbone, and requires minimal GPU memory (Table 7).

Table 7: GPU Memory. Based on AttMPTI Setting. ($S_0$)

| Method | 2-way 1-shot | 2-way 5-shot |
|---|---|---|
| AttMPTI | ~10GB | ~29GB |
| DSG (ours) | ~11GB | ~30GB |

# D More Ablation Studies

## D.1 Feature Decoupling Verification

We calculated the CKA (Kornblith et al., 2019) and RSA (Kriegeskorte et al., 2008) values between semantic (TMSA) and geometric features (SRI) under different datasets and settings. As shown in Table 8, the CKA are extremely low (all <0.01), and the RSA are close to 0. These results are descriptive evidence of representation differences, not proof of statistical independence or complete disentanglement. The CKA/RSA protocol is given in App. F.4; additional functional comparisons appear in App. D.2.

Table 8: CKA and RSA Analysis. ($S_0$)

| | **S3DIS** | | **ScanNet** | |
|---|---|---|---|---|
| | 2-way 1-shot | 2-way 5-shot | 2-way 1-shot | 2-way 5-shot |
| CKA | $6.40 \times 10^{-3}$ | $8.90 \times 10^{-3}$ | $2.40 \times 10^{-3}$ | $7.60 \times 10^{-3}$ |
| RSA | $-1.35 \times 10^{-2}$ | $-4.09 \times 10^{-2}$ | $-3.90 \times 10^{-2}$ | $-1.60 \times 10^{-3}$ |

## D.2 Additional Decoupling Comparison

Table 9 reports additional functional and representation-level comparisons with a no-decoupling variant and a coupling proxy. Full DSG achieves the highest mIoU. Low representation similarity alone does not imply better disentanglement; these scores are considered alongside segmentation performance.

Table 9: Additional comparison of semantic–geometric representations.

| Method | mIoU | CKA | RSA |
|---|---|---|---|
| No decouple | 40.59 | 0.010973 | -0.002964 |
| Coupling proxy | 42.66 | 0.002318 | -0.074871 |
| Full DSG | **50.22** | 0.006400 | -0.013500 |

## D.3 More TMSA and SRI Ablation

We comprehensively extended component analysis on the S3DIS and ScanNet datasets for both 1-shot and 5-shot (2-way) settings. As shown in Table 10, the relative contributions of TMSA and SRI remained highly consistent across different datasets and settings. The TMSA module consistently established a strong multimodal semantic baseline; subsequently, the SRI module consistently delivered the final performance leap as a powerful pushover geometric regularizer. Both modules were indispensable in all test scenarios.

Table 10: Ablation study of TMSA and SRI. ($S_0$)

| | S3DIS | | ScanNet | | Mean |
|---|---|---|---|---|---|
| | 2-way 1-shot | 2-way 5-shot | 2-way 1-shot | 2-way 5-shot | |
| Baseline | 38.63 | 42.51 | 39.56 | 46.53 | 41.81 |
| +TMSA | 42.90 | 44.90 | 47.33 | 53.57 | 47.18 |
| +TMSA+SRI | **50.22** | **48.99** | **50.58** | **57.67** | **51.87** |

### D.4 Impact of the TMSA Depth

As shown in Table 11, the model achieves peak performance with a single cross-attention layer ($K = 1$), attaining an mIoU of 50.22%. Increasing the depth to two or three layers leads to a notable degradation in performance. Overall, this trend indicates that shallow interaction mechanisms are more effective in few-shot settings.

Table 11: Impact of the TMSA Depth.

| $K$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| mIoU | **50.22** | 47.38 | 47.26 | 49.44 | 48.75 |

### D.5 Sensitivity Analysis

As shown in Tables 12a and 12b, a detailed grid evaluation of $N_p, N_a$ was performed. DSG was relatively stable over a wide range. When the parameter value is 40, a performance decline is generally observed, which precisely triggers semantic fragmentation or excessive manifold segmentation, basically in line with mathematical expectations. To ensure robustness, we did not tune the parameters for the dataset, adopted the $N_p = 50, N_a = 50$.

The S3DIS results in Tables 12a and 12b show the impact of $N_p$ and $N_a$. The parameter $N_p$ controls the Gaussian mixture model's ability to capture intra-class semantic variations. Increasing $N_p$ does not lead to monotonic accuracy gains. Similarly, Na controls the density of unsupervised Voronoi mosaicking. The highest reported value is at $N_a = 50$, with mIoU 50.22%. Together with the timing analysis in App. G.4, the default moderate budget provides a favorable accuracy–efficiency trade-off.

Table 12: Sensitivity Analysis. (2-way 1-shot $S_0$)

(a) Sensitivity to $N_p$

| $N_p$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|
| S3DIS | 47.47 | 49.47 | **50.33** | 47.66 | 50.22 | 46.63 | 47.43 | 47.00 | 49.72 |
| ScanNet | 49.73 | 49.27 | 49.06 | 49.79 | **50.58** | 49.04 | 50.05 | 50.20 | 49.48 |
| Average | 48.60 | 49.37 | 49.70 | 48.73 | **50.40** | 47.84 | 48.74 | 48.60 | 49.60 |

(b) Sensitivity to $N_a$

| $N_a$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|
| S3DIS | 46.04 | 47.93 | 48.95 | 45.25 | **50.22** | 50.19 | 47.33 | 47.66 | 49.27 |
| ScanNet | 48.85 | 50.14 | 49.20 | 47.87 | **50.58** | 49.47 | 49.68 | 45.47 | 47.17 |
| Average | 47.45 | 49.04 | 49.08 | 46.56 | **50.40** | 49.83 | 48.51 | 46.57 | 48.22 |

### D.6 Semantic Prototype Strategies

Table 13 compares the reported semantic prototype strategies. The additional multi-prototype averaging result is 48.59 on S3DIS. The contribution of DSG lies in semantic–geometric inference rather than a new clustering update. Under identical class masks, FPS initialization, feature spaces, and




Figure 6: Extended qualitative comparisons on the S3DIS dataset (2-way 1-shot).

a single update cycle, Lloyd k-means and our Hard-EM implementation are numerically equivalent (App. F.1).

Table 13: Semantic prototype strategies (2-way 1-shot $S_0$). A dash denotes a result not reported for that strategy.

| Method | S3DIS | ScanNet |
|---|---|---|
| Multi-prototype Average | 48.59 | – |
| k-means | 46.77 | 47.93 |
| DSG (Hard-EM) | **50.22** | **50.58** |

# E More Visual Experimental Analysis

## E.1 Extended Qualitative Results

To further assess the robustness of DSG across a broader range of scene categories and geometric complexities, we present extended visualizations on the S3DIS dataset in Figure 6. These examples

highlight the limitations of the coupled baseline (MM-FSS) and demonstrate how our probabilistic decomposition strategy effectively mitigates common failure modes.

**Mitigating boundary leakage.** A persistent challenge in point cloud segmentation is distinguishing geometrically similar objects from their surroundings, such as a board mounted on a wall. In row 1, DSG reduces background leakage compared with MM-FSS. However, DSG misses the small board region in row 6, illustrating a failure case under limited foreground coverage. In contrast, DSG, empowered by the SRI module, implicitly captures fine-grained topological boundaries through unsupervised manifold anchors, resulting in clear boundary segmentation that closely conforms to the physical extent of the object.

**Maintaining structural fidelity for complex objects.** For categories with high geometric complexity and substantial intra-class variance, such as chairs, baseline methods often yield incomplete predictions. As illustrated in line 2, MM-FSS fails to capture the internal lattice structure, resulting in fragmented clusters. In contrast, our DSG explicitly models semantic and geometric subspaces, enabling it to preserve structural continuity even when semantic features are noisy.

### E.2 Component-wise Visual Analysis

To gain deeper insight into the specific contributions of our decomposed probability flow, we present a stepwise visual ablation analysis in Figure 7. This visualization empirically validates the decoupling between semantic likelihood estimation and geometric prior modeling.

**Semantic disambiguation via TMSA.** The Baseline column indicates that the coupled feature space suffers from severe semantic ambiguity, manifested by widespread false positives (e.g., in rows 2, 4, 5, 6, and 8, background walls are erroneously segmented as bookcases) and false negatives (e.g., in rows 3 and 5, the true extent of the bookcase is not correctly identified and is instead mislabeled as background). After integrating the TMSA module, a pronounced "cleaning" effect is observed: by maximizing semantic likelihood, the model effectively suppresses these semantic artifacts. Nevertheless, predictions in the +TMSA column remain relatively coarse and blocky, lacking fine-grained internal structural detail.

**Geometric correction via SRI.** As shown in the +SRI column, the introduction of a geometric prior serves as a topological corrector. This effect is most evident for the bookcase in row 3, where SRI successfully recovers internal geometry lost in the previous stage. In row 8, adding SRI reduces false-positive predictions around the bookcase. Despite these improvements, there remains room for further refinement.

### E.3 Future Work

We proposed DSG based on benchmarked COSeg, achieves substantial improvements in evaluation metrics. However, in practical settings, the scenes remain densely segmented, and current 2-way 5-shot configurations require nearly 80 GB of GPU memory (MM-FSS setting), resulting in considerable computational overhead. This limitation renders few-shot training infeasible at higher settings, such as 3-way 5-shot. An important direction for future work is to substantially reduce this overhead, thereby enabling training under more challenging configurations.




Figure 7: Step-wise visual ablation breakdown on S3DIS (2-way 1-shot).

# F Additional Implementation and Evaluation Details

## F.1 Single-step Lloyd/Hard-EM Implementation

The probabilistic formulation specifies the statistical object being estimated: a class-conditional feature distribution with multiple latent intra-class modes. It does not make the optimization stochastic. Under equal mixture weights, shared isotropic covariance, and the zero-variance limit, the Hard-EM E-step assigns each feature to its nearest prototype and the M-step computes the mean of the assigned features. With fixed FPS initialization, these operations are deterministic. Our implementation performs exactly one assignment-and-mean update cycle, numerically equivalent to one Lloyd update under the same class masks, initialization, feature spaces, and update count. We do not introduce a new clustering update rule.

**Shot-wise semantic prototypes.** For the tested configurations, $N_p$ is divisible by the support-shot count $K$. Each support shot independently extracts $N_p^{\text{shot}} = N_p/K$ prototypes within each class mask; concatenating the shot-wise prototypes gives $N_p$ prototypes per class. For class $c$ and shot $r$, FPS is performed on the corresponding 3D coordinates:

$$I^{c,r} = \text{FPS}(\mathbf{x}_s^{c,r}, N_p^{\text{shot}}), \qquad \boldsymbol{\mu}_l^{b,c,r,(0)} = \mathbf{F}_s^{b,c,r}[I_l^{c,r}], \quad b \in \{\text{i}, \text{u}\}. \tag{27}$$

The IF and UF branches share the FPS indices but independently compute distances, assignments, and means in their respective feature spaces. Suppressing the branch, class, and shot indices, the update is

$$z_{j,l} = \mathbf{1}\left[l = \operatorname*{argmin}_m \|\mathbf{f}_j - \boldsymbol{\mu}_m^{(0)}\|_2^2\right], \qquad \boldsymbol{\mu}_l^{(1)} = \frac{\sum_j z_{j,l}\mathbf{f}_j}{\sum_j z_{j,l}}. \tag{28}$$

For an empty component, the FPS-initialized feature is retained. No additional explicit $L_2$ normalization is applied before assignment. Prototype estimation is performed online in each episode. Gradients do not propagate through FPS or the $\operatorname{argmin}$ decisions, but can propagate through the mean update to the assigned features.

The discrete component indices are not semantic labels. Instead, cosine affinities between each query point and all prototypes of a class are retained and aggregated through learnable mappings. The procedure is therefore FPS-initialized single-step Lloyd/Hard-EM refinement for class-conditional prototype estimation. Finite-variance Soft-EM, non-uniform mixture weights, and component-specific covariances are not implemented in this work.

**Query structural anchors.** Equation 9 is the classical squared-error vector-quantization objective. We initialize anchors with 3D FPS and independently perform one nearest-anchor assignment and one mean update in the IF and UF feature spaces. No query masks or semantic labels are used. The resulting anchors construct continuous affinities

$$\mathbf{G}^b = \text{sim}(\mathbf{F}_q^b, \mathbf{A}^{b,(1)}), \qquad b \in \{\text{i}, \text{u}\}. \tag{29}$$

Hard assignments are used only to update the anchors; they are not segmentation predictions. Geometric clusters need not correspond one-to-one with semantic categories. Point-to-anchor affinities provide category-independent query structure, which is combined with support-derived semantic evidence for subsequent inference. This is FPS-initialized single-step Lloyd refinement for query-manifold anchor extraction, rather than an alternative optimization algorithm to k-means.

## F.2 Dataset and Episode Protocol

Following the COSeg and MM-FSS protocols, scenes are divided into $1\,\text{m} \times 1\,\text{m}$ blocks. Each block belongs to a single scene, is voxelized with a $0.02\,\text{m}$ grid using uniform sampling, and retains at most 20,480 points. We use 271 rooms from S3DIS and 1,613 scenes from ScanNet following the standard dataset splits. S3DIS has 12 evaluated categories after excluding clutter, with six test classes per split. ScanNet has 20 categories, with ten test classes per split. The test classes are unseen during training; their names are listed in Table 14.

During training, 400 episodes are randomly sampled per epoch. Evaluation uses 1,000 episodes per class in the 1-way setting and 100 episodes per class combination in the 2-way setting. For example,

Table 14: Test classes in the two splits of S3DIS and ScanNet.

| Dataset | $S^0$ | $S^1$ |
|---|---|---|
| S3DIS | beam, board, bookcase, ceiling, chair, column | door, floor, sofa, table, wall, window |
| ScanNet | bathtub, bed, bookshelf, cabinet, chair, counter, curtain, desk, door, floor | otherfurniture, picture, refrigerator, shower curtain, sink, sofa, table, toilet, wall, window |

a S3DIS split has $\binom{6}{2} = 15$ two-class combinations, giving $15 \times 100 = 1{,}500$ episodes. We calculate IoU for each target class and average over the test classes to obtain mIoU. Dataset-level results are averaged over the predefined test splits.

### F.3 Baseline Sources and Evaluation Settings

Table 15 summarizes the backbone, VLM, and episode protocol for the main benchmark comparisons. Results quoted from previous papers are distinguished from our additional checkpoint re-evaluations. For MM-FSS and DA-FSS, the additional evaluation uses their released checkpoints under the same episode-sampling protocol as DSG, with consistent backbone, modalities, and preprocessing. This checkpoint evaluation is separate from the SRI-transfer experiment in App. G.2, in which the augmented methods are retrained.

Table 15: Protocol comparison for the main benchmark. † denotes the multimodal COSeg variant.

| Method | Backbone | VLM | Episode protocol |
|---|---|---|---|
| AttMPTI | Stratified Transformer | No | Same |
| QGE | Stratified Transformer | No | Same |
| QGPA | Stratified Transformer | No | Same |
| COSeg | Stratified Transformer | No | Same |
| COSeg† | Stratified Transformer | CLIP | Same |
| MM-FSS | Stratified Transformer | CLIP | Same |
| DA-FSS | Stratified Transformer | CLIP | Same |
| DSG | Stratified Transformer | CLIP | Same |

### F.4 Representation Similarity Protocol

We extract the semantic representations generated by TMSA and the geometric representations modeled by SRI, denoted by $\mathbf{Z}_s$ and $\mathbf{Z}_g$, across multiple evaluation episodes. Linear CKA is computed as

$$\mathrm{CKA}(\mathbf{Z}_s, \mathbf{Z}_g) = \frac{\|\mathbf{Z}_s^\top \mathbf{Z}_g\|_F^2}{\|\mathbf{Z}_s^\top \mathbf{Z}_s\|_F \, \|\mathbf{Z}_g^\top \mathbf{Z}_g\|_F}. \tag{30}$$

For RSA, cosine-similarity matrices $\mathbf{M}_s$ and $\mathbf{M}_g$ are constructed from normalized features. We compute the Pearson correlation between their upper-triangular elements:

$$\mathrm{RSA}(\mathbf{Z}_s, \mathbf{Z}_g) = \rho\big(\mathrm{vec}(\mathbf{M}_s^{\mathrm{upper}}), \mathrm{vec}(\mathbf{M}_g^{\mathrm{upper}})\big). \tag{31}$$

These measures provide representation-level evidence about redundancy and complementary information, not a strict proof of independence. Unrelated representations may also have low similarity; lower similarity alone does not imply better disentanglement.

### F.5 NOTATION

Table 16: Notation used in the method and additional evaluation.

| Symbol | Meaning |
|---|---|
| $N$, $K$ | Number of target classes and support shots in an $N$-way $K$-shot episode. |
| $M_s$, $M_q$ | Numbers of support and query points. |
| $N_p$, $N_a$ | Numbers of semantic prototypes per class and query structural anchors; both are 50 by default. |
| $\phi$ | Parameters of the learned encoder. |
| $\mathbf{S}_k$ | Semantic affinity at the $k$-th TMSA refinement step. |
| $K$ (TMSA depth) | Number of TMSA refinement steps in the depth ablation, distinct from the shot count. |
| $\mathbf{G}_{\text{raw}}$ | Sum of IF and UF point-to-anchor affinity matrices. |
| $\tau$ | Temperature scale in the final cosine-softmax classifier. |
| $t$, $l$ | Update index and prototype/anchor index in the assignment equations. |
| $k_{\text{nn}}$ | Neighborhood size used only in boundary evaluation. |

## G ADDITIONAL EXPERIMENTAL RESULTS

### G.1 STABILITY ACROSS INDEPENDENT EPISODE SETS

We keep the trained models fixed and evaluate DSG, MM-FSS, and DA-FSS on three independent test episode sets generated with different seeds under the original protocol. Tables 17 and 18 report mean mIoU and standard deviation across these episode sets, with results averaged over $S^0$ and $S^1$. These statistics describe evaluation-episode sampling, rather than variation across independently trained models. DSG obtains the highest mean in all eight dataset/setting combinations.

Table 17: S3DIS mIoU (mean $\pm$ std) over three independent episode sets, averaged over $S^0$ and $S^1$.

| Method | 1-way 1-shot | 1-way 5-shot | 2-way 1-shot | 2-way 5-shot |
|---|---|---|---|---|
| DA-FSS | $52.74 \pm 0.43$ | $50.32 \pm 0.27$ | $43.73 \pm 0.23$ | $46.66 \pm 0.08$ |
| MM-FSS | $51.72 \pm 0.36$ | $53.87 \pm 0.07$ | $44.04 \pm 0.29$ | $50.06 \pm 0.58$ |
| DSG | $\mathbf{56.26} \pm 0.18$ | $\mathbf{60.30} \pm 1.92$ | $\mathbf{51.61} \pm 0.65$ | $\mathbf{53.56} \pm 2.47$ |

Table 18: ScanNet mIoU (mean $\pm$ std) over three independent episode sets, averaged over $S^0$ and $S^1$.

| Method | 1-way 1-shot | 1-way 5-shot | 2-way 1-shot | 2-way 5-shot |
|---|---|---|---|---|
| DA-FSS | $45.53 \pm 0.16$ | $50.45 \pm 0.33$ | $39.27 \pm 0.10$ | $44.85 \pm 0.27$ |
| MM-FSS | $44.59 \pm 0.15$ | $50.30 \pm 0.34$ | $39.31 \pm 0.12$ | $44.15 \pm 0.36$ |
| DSG | $\mathbf{48.98} \pm 2.78$ | $\mathbf{55.01} \pm 2.50$ | $\mathbf{47.18} \pm 2.57$ | $\mathbf{51.98} \pm 2.74$ |

### G.2 TRANSFER OF SRI TO EXISTING METHODS

We integrate SRI into MM-FSS and DA-FSS. Each baseline supplies its query semantic affinity as the semantic input and its query feature map for estimating $N_a = 50$ structural anchors. No query annotations are used. The augmented methods are retrained with their original training settings, and all variants are evaluated on the same S3DIS $S^0$, 2-way 1-shot episode set. Table 19 shows gains of 6.61 and 1.91 mIoU points, respectively, supporting the applicability of SRI beyond DSG in this setting.

### G.3 BOUNDARY EVALUATION AND FAILURE CASES

A query point is defined as a boundary point if at least one of its $k_{\text{nn}}$ nearest neighbors has a different ground-truth semantic label. Boundary mIoU is computed only on the resulting boundary points.

Table 19: SRI transfer on S3DIS $S^0$, 2-way 1-shot.

| Method | Baseline | +SRI | Improvement |
|---|---|---|---|
| MM-FSS | 41.98 | 48.59 | +6.61 |
| DA-FSS | 42.07 | 43.98 | +1.91 |

The same evaluation protocol is applied to all compared methods. Table 20 reports results for $k_{\mathrm{nn}} \in \{4, 8, 16, 32, 64\}$, using a separate symbol to distinguish neighborhood size from support-shot count.

Table 20: Boundary-aware evaluation under five neighborhood sizes.

| | | Boundary mIoU | | | | |
|---|---|---|---|---|---|---|
| Method | mIoU | $k_{\mathrm{nn}} = 4$ | $k_{\mathrm{nn}} = 8$ | $k_{\mathrm{nn}} = 16$ | $k_{\mathrm{nn}} = 32$ | $k_{\mathrm{nn}} = 64$ |
| DA-FSS | 40.97 | 32.79 | 33.58 | 35.04 | 36.40 | 38.35 |
| MM-FSS | 41.86 | 34.24 | 35.03 | 36.46 | 38.19 | 40.04 |
| DSG | **50.22** | **39.12** | **39.99** | **41.54** | **43.18** | **45.64** |

DSG achieves higher boundary mIoU than both baselines for all five neighborhood sizes. For $k_{\mathrm{nn}} = 4$, it obtains 39.12 compared with 34.24 for MM-FSS and 32.79 for DA-FSS. Boundary regions remain more challenging than the overall scene. As illustrated by the qualitative results in Figures 6 and 7, failures can still occur in sparse regions, at ambiguous boundaries, and between adjacent objects with similar geometric structures. Predicted boundaries may shift toward object interiors, causing local under-segmentation. Geometric regularization can therefore be less reliable when semantic and geometric cues conflict.

### G.4 FPS Efficiency and Prototype Budget

For an input with $M$ points, vanilla FPS takes $O(M^2)$ operations when selecting all points. Selecting a fixed budget of $N_p$ initial centroids instead costs approximately $O(MN_p)$. We use $N_p = 50$ by default. Table 21 measures FPS and end-to-end inference times on an RTX A6000 for inputs up to 81,920 points. The FPS overhead is small relative to the measured inference time.

Table 21: Input-size scaling on an RTX A6000.

| Input points | $N_p$ | FPS (ms) | Inference (s) |
|---|---|---|---|
| 20,480 | 50 | 2.19 | 0.92 |
| 40,960 | 50 | 3.05 | 3.17 |
| 81,920 | 50 | 3.17 | 3.60 |

We also vary $N_p$ from 10 to 90 with 40,960 input points (Table 22). FPS latency and overall inference time generally increase with the sampling budget, although the measurements are not strictly monotonic. A larger prototype set also increases subsequent point-to-prototype affinity computation. Together with the accuracy sensitivity results in Tables 12a and 12b, these measurements support using a moderate budget for a favorable accuracy–efficiency trade-off.

Table 22: Prototype-budget timing with 40,960 input points.

| $N_p$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|
| FPS (ms) | 2.79 | 3.02 | 2.93 | 3.01 | 3.05 | 3.19 | 3.33 | 3.45 | 3.49 |
| Inference (s) | 2.97 | 2.78 | 3.09 | 3.14 | 3.17 | 3.24 | 3.29 | 3.34 | 3.40 |