IFOT-SSEG: INDUCED FACTOR OPTIMAL TRANSPORT FOR UNSUPERVISED SEMANTIC SEGMENTATION OF 3D POINT CLOUDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised 3D point cloud segmentation often relies on clustering learned region features, but mixed latent factors and unreliable fragment-to-primitive pseudo-label assignment jointly limit the quality of pseudo supervision. To address this problem, we propose a latent factor-driven global pseudo-label inference framework that first recovers semantic-relevant latent factors from neural region features and then performs globally competitive fragment-to-primitive assignment to generate reliable pseudo-labels. Specifically, Bayesian Nonparametric Latent Factor Induction (BN-LFI) models neural region features as observations from an open-ended latent factor space and combines ICA-style factor construction with variational Indian Buffet Process inference to adaptively recover active factors underlying the current scene semantics. These factors are further calibrated using unsupervised quality signals, yielding a more reliable latent representation for subsequent pseudo-label inference. Based on the induced representation, Global Margin-Aware Optimal Transport (GM-OT) formulates fragment-to-primitive-slot matching as a globally competitive transport problem, suppressing dominant primitive absorption and slot redundancy while producing more stable hard pseudo-labels. The framework operates only during pseudo-label construction, introduces no additional inference overhead, and outperforms existing unsupervised 3D point cloud segmentation methods on ScanNet, S3DIS, and nuScenes.

1 INTRODUCTION

3D point cloud semantic segmentation is a fundamental task in 3D scene understanding, providing essential environmental semantics for robot perception, embodied intelligence, and autonomous driving. Recent advances in deep neural networks and 3D sensing have substantially improved supervised 3D segmentation (Qi et al., 2017a;b; Graham et al., 2018), yet these methods rely heavily on large-scale point-wise annotations. Since 3D point clouds are sparse, irregular, and structurally complex, high-quality semantic annotation is expensive and difficult to scale with the rapidly growing volume of real-world data (Dai et al., 2017). This has motivated increasing interest in unsupervised 3D segmentation, which aims to discover semantic structures directly from unlabeled point clouds (Zhang et al., 2023; Chen et al., 2023; Liu et al., 2024; Zhang et al., 2025b; Song et al., 2026). Early approaches directly clustered point- or region-level features (Zhang et al., 2023; Chen et al., 2023; Liu et al., 2024), while subsequent methods introduced superpoint aggregation to improve local structural consistency (Zhang et al., 2023; 2025b) or transferred semantic priors from 2D foundation models to strengthen 3D representations (Chen et al., 2023). More recent studies further improve pseudo-label quality through region grouping, local–global consistency, and continuous representation learning (Zhang et al., 2025b; Song et al., 2026). Despite these advances, existing methods still largely construct pseudo supervision by clustering learned region features and progressively merging fine-grained fragments according to prototype similarity.

The reliability of this pseudo-label construction pipeline is constrained by the interaction between region representation and fragment assignment. Existing methods typically discover semantic structures directly from fixed-dimensional neural-region features (Zhang et al., 2023; Chen et al., 2023; Zhang et al., 2025b). However, such features contain responses associated with semantics, geometry, local structures, and other underlying factors, while only part of them may be useful for

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

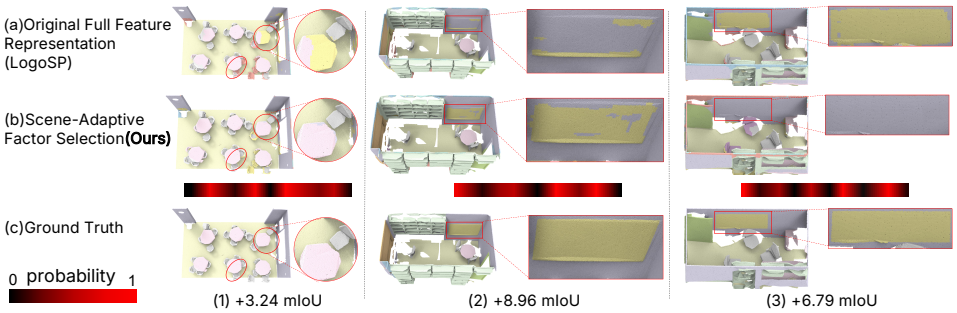


Figure 1: Comparison between the original full-feature representation (Zhang et al., 2025b) and scene-adaptive latent factor selection. Scenes (1) and (2) illustrate factor selection across different scene contents. Scenes (2) and (3) show boards in different scene contexts, motivating the selection of different discriminative factors even for the same semantic category. BN-LFI infers different subsets of latent factors across scenes, suppresses redundant responses, and improves scene-level mIoU. The bars indicate posterior factor inclusion probabilities.

semantic discrimination in a particular scene. Directly clustering the full representation can therefore introduce redundant responses into prototype estimation, making the discovered semantic structures less reliable, as illustrated in Fig. 1. This representation uncertainty is further amplified during pseudo-label assignment. After obtaining fine-grained fragments, existing methods commonly merge them into primitive prototypes according to local feature similarity (Zhang et al., 2023; Chen et al., 2023; Zhang et al., 2025b). Since these prototypes are estimated from evolving pseudo-assignments rather than ground-truth class centers, early representation and assignment errors can progressively bias subsequent prototype updates. Dominant primitives can consequently attract additional ambiguous or fine-grained fragments and further reinforce their prototypes, whereas weak primitive slots receive insufficient support and may gradually become redundant. Moreover, independent local matching does not explicitly model the competition of multiple candidate primitives for the same fragment, allowing ambiguous low-margin correspondences to be directly converted into hard pseudo-labels. These observations reveal a coupled challenge: reliable pseudo-label construction requires both identifying the latent factors that truly support semantic discrimination and using the resulting representation to resolve fragment-to-primitive correspondences under global competition.

To address this challenge, we propose a *latent factor-driven framework for global pseudo-label inference* that progressively connects latent semantic factor induction with globally competitive fragment-to-primitive assignment. We first introduce Bayesian Nonparametric Latent Factor Induction (BN-LFI), which treats neural-region features as finite observations from an open-ended latent factor space. BN-LFI constructs weakly correlated candidate factors through ICA-style factor induction and employs a variational Indian Buffet Process to adaptively infer the active factors and their effective number for different regions, allowing the representation to adjust to scene-dependent semantic structures. We further calibrate the induced factors using unsupervised teacher-quality signals, where counterfactual enhancement and suppression estimate their contributions to pseudo-label quality and suppress redundant responses. The resulting latent representation then serves as the basis for Global Margin-Aware Optimal Transport (GM-OT), which formulates spectral-fragment-to-primitive-slot merging as a globally coupled transport problem rather than independent local matching. GM-OT jointly optimizes fragment-to-slot correspondences while accounting for primitive capacity, induced feature distance, fragment reliability, structural risk, and competitive margin, thereby limiting the unchecked expansion of dominant primitives and discouraging ambiguous assignments. The global transport solution is subsequently used to update primitive prototypes and generate more stable hard pseudo-labels for training the 3D segmentation network. Both BN-LFI and GM-OT operate only during pseudo-teacher construction and therefore introduce no additional inference overhead.

Our main contributions are summarized as follows:

- We propose a latent factor-driven framework for global pseudo-label inference that improves pseudo-label reliability in unsupervised 3D point cloud segmentation through latent semantic representation and globally competitive assignment.

- We introduce Bayesian Nonparametric Latent Factor Induction (BN-LFI), which combines ICA-style factor construction, variational IBP-based sparse inference, and unsupervised causal calibration to adaptively identify and select active latent factors contributing to the current semantic structure.
- We introduce Global Margin-Aware Optimal Transport (GM-OT), which formulates fragment-to-primitive slot merge-back as globally competitive assignment under capacity constraints. By jointly considering fragment reliability, structural risk, and competitive margins, it mitigates over-absorption by dominant primitives and slot redundancy.
- Extensive experiments on ScanNet, S3DIS, and nuScenes demonstrate that our method outperforms existing unsupervised 3D point cloud segmentation approaches and validate the effectiveness of both core modules.

2 Related Work

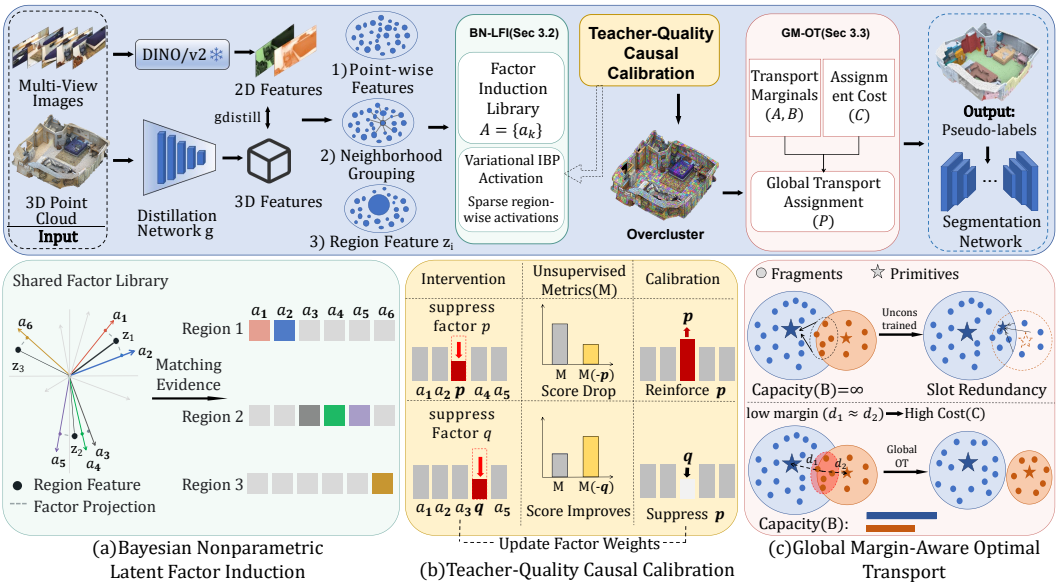
Unsupervised 3D Segmentation: Unsupervised 3D semantic segmentation aims (Zhang et al., 2023; Chen et al., 2023; Liu et al., 2024; Zhang et al., 2025b) to discover semantic structures from point clouds without manual annotations. Clustering-based 3D methods construct pseudo-labels from point- or region-level representations (Zhang et al., 2023; Chen et al., 2023). GrowSP (Zhang et al., 2023) progressively groups superpoints into semantic primitives, PointDC (Chen et al., 2023) combines cross-modal distillation with super-voxel clustering, and LogoSP (Zhang et al., 2025b) exploits local–global superpoint patterns in the spectral domain. PointGS (Song et al., 2026) further introduces a continuous Gaussian representation to improve cross-modal semantic transfer. Despite this progress, semantic discovery in these pipelines remains largely dependent on similarity-based clustering or prototype assignment, making pseudo labels sensitive to feature ambiguity, structural imbalance, and dominant regional patterns.

Foundation-Model Priors for 3D Representation: Large-scale 2D foundation models (Caron et al., 2021; Oquab et al., 2024) provide rich semantic priors for 3D scene understanding (Chen et al., 2023; Zhang et al., 2025b). A common strategy projects multi-view image features into 3D space and distills them into a point-cloud backbone (Chen et al., 2023; Peng et al., 2023; Zhang et al., 2025b). PointDC (Chen et al., 2023) transfers self-supervised visual features before super-voxel clustering, while LogoSP uses DINO/DINOv2-style representations (Caron et al., 2021; Oquab et al., 2024; Zhang et al., 2025b) to initialize superpoint grouping. Although such priors improve semantic separability, occlusion, sparse sampling, viewpoint inconsistency, and the image-to-point-cloud domain gap can still introduce ambiguous region features. Moreover, foundation-model representations are general-purpose and contain diverse semantic factors, where only a subset may be relevant to specific scenes. Therefore, reliable pseudo-label construction requires further representation refinement beyond directly using distilled features.

Adaptive Feature Refinement for Semantic Segmentation: Adaptive feature selection and re-weighting have been widely explored to improve discriminative representations for semantic segmentation. Early work learns to adaptively weight multi-scale features for pixel-wise prediction (Chen et al., 2016), while EncNet (Zhang et al., 2018) uses scene-level semantic context to selectively emphasize class-dependent feature maps. DFN (Yu et al., 2018) further introduces channel attention to enhance discriminative features and suppress less informative responses, and DANet (Fu et al., 2019) models spatial and channel dependencies to selectively aggregate semantically relevant features. Similar ideas have also been explored in 3D segmentation: S3Net (Cheng et al., 2021a) emphasizes informative responses through intra- and inter-channel attention, while AF2-S3Net (Cheng et al., 2021b) explicitly performs adaptive feature selection with feature map re-weighting. However, these methods mainly select or re-weight features within predefined representation spaces and are optimized with semantic supervision. They do not explicitly discover the latent factors underlying high-dimensional representations or adapt their factor structure to different scenes, leaving scene-dependent factor selection for unsupervised pseudo-label construction largely unexplored.

3 Methodology

Given an unlabeled point cloud $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N_p}$, our framework constructs point-level pseudo-labels from distilled point features, as illustrated in Fig. 2. The upper pipeline consists of region formation,



(a)Bayesian Nonparametric Latent Factor Induction (b)Teacher-Quality Causal Calibration (c)Global Margin-Aware Optimal Transport

Figure 2: **Overview**. The upper row shows the complete pseudo-label construction pipeline. Distilled point-wise features are grouped into neural regions, from which BN-LFI constructs a latent factor library, infers sparse region-wise activations, and calibrates them using teacher-quality feedback. The induced representations are overclustered into candidate fragments, which GM-OT globally assigns to primitive slots according to transport costs and capacity constraints. The resulting pseudo-labels supervise a standard 3D segmentation network. The lower panels detail **(a) Latent Factor Induction**, **(b) Causal Calibration**, and **(c) Global Optimal Transport**.

latent factor induction, candidate fragment discovery, and global transport assignment:

$$\mathbf{Z} = \mathcal{R}(E_\psi(\mathcal{P})), \quad \mathbf{Z}^{\text{final}} = \mathcal{I}_{\text{BN-LFI}}(\mathbf{Z}), \quad \mathcal{F} = \mathcal{C}_{\text{spec}}(\mathbf{Z}^{\text{final}}), \quad \widehat{\mathbf{y}}^F = \mathcal{T}_{\text{GM-OT}}(\mathcal{F}). \tag{1}$$

Here, E_ψ extracts distilled point-wise features and $\mathcal{R}$ groups them into neural regions. BN-LFI constructs a shared latent factor library, infers sparse region-wise activations, and calibrates them using teacher-quality feedback, as detailed in the lower-left and middle panels of Fig. 2. The calibrated regions are overclustered into candidate fragments, which GM-OT jointly assigns to unordered primitive slots under transport costs and capacity constraints. The fragment assignments $\widehat{\mathbf{y}}^F$ are propagated to their constituent points to produce hard pseudo-labels for training a standard 3D segmentation network.

3.1 Neural-Region Feature Observations

Following prior 2D-to-3D distillation methods (Chen et al., 2023; Zhang et al., 2025b), a pretrained backbone E_ψ extracts a feature $\mathbf{f}_i \in \mathbb{R}^D$ for each point $\mathbf{p}_i$. Initial superpoints (Landrieu & Simonovsky, 2018) are further grouped by feature similarity into neural regions (Zhang et al., 2023). For region $\mathcal{R}_n$, its representation and the stacked region-feature matrix are

$$\mathbf{z}_n = \frac{1}{|\mathcal{R}_n|} \sum\nolimits_{\mathbf{p}_i \in \mathcal{R}_n} \mathbf{f}_i, \qquad \mathbf{Z} = [\mathbf{z}_n]_{n=1}^N \in \mathbb{R}^{N \times D}. \tag{2}$$

The resulting $\mathbf{Z}$ provides compact observations for pseudo-label construction, but its channels remain mixtures of multiple latent semantic-geometric factors.

3.2 Bayesian Nonparametric Latent Factor Induction

Although neural-region aggregation reduces point-level noise, each channel of $\mathbf{Z}$ may still mix geometry, boundaries, density, viewpoint, and semantic responses, while one discriminative factor may span multiple channels. Restricting candidate factors to the original channels or a predefined finite partition therefore assumes that all segmentation-relevant factors have been enumerated. Since

different regions may require different factor combinations and numbers, BN-LFI instead introduces an open candidate factor space from which each region activates a finite, data-adaptive subset.

We model the open-ended candidate factor space as a countably infinite library $\mathcal{A} = \{\mathbf{a}_k\}_{k=1}^{\infty} \subset \mathbb{R}^D$ (Griffiths & Ghahramani, 2011). Each vector $\mathbf{a}_k$ represents a candidate semantic-geometric factor direction in the observed feature space. The library is shared across the unlabeled training scenes, whereas factor activation is specific to each neural region. Accordingly, the feature of region i is approximated by

$$\mathbf{z}_i \approx \sum\nolimits_{k=1}^{\infty} b_{ik} s_{ik} \mathbf{a}_k + \boldsymbol{\epsilon}_i, \qquad b_{ik} \in \{0, 1\}, \tag{3}$$

where b_{ik} indicates whether region i activates factor k, s_{ik} is its continuous response, and $\boldsymbol{\epsilon}_i$ captures residual variations. Thus, different regions can activate different finite subsets of the same open candidate library.

For tractable inference, we use an overcomplete truncation (Doshi et al., 2009) $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{K_{\max}}]^\top \in \mathbb{R}^{K_{\max} \times D}$ with $K_{\max} > D$. Here, $K_{\max}$ is only a computational upper bound; the effective number of factors used by each region is determined by the variational IBP posterior.

Data-driven latent factor library construction: We whiten the neural-region features and learn the truncated factor library $\mathbf{A}$ through overcomplete ICA-style factor construction (Hyvärinen & Oja, 2000; Knowles & Ghahramani, 2007). The resulting directions capture recurring semantic-geometric variations in the unlabeled feature distribution, while encouraging sparse region responses and reducing inter-factor redundancy. Because ICA-style factors are permutation-indeterminate whereas the stick-breaking prior is size-biased (Teh et al., 2007), we order the candidate factors by their empirical sparse-response mass before variational inference. The resulting response-aligned library provides data-supported candidate directions for region-specific IBP activation.

Variational IBP factor activation: The response-aligned library $\mathbf{A}$ provides candidate factor directions, but each neural region may require a different sparse subset. We therefore place a stick-breaking IBP prior on the region-factor activations (Teh et al., 2007), where $v_k \sim \mathrm{Beta}(\alpha, 1)$, $\pi_k = \prod_{\ell=1}^{k} v_\ell$, and $b_{ik} \sim \mathrm{Bernoulli}(\pi_k)$. Here, b_{ik} indicates whether region i activates factor k, while the decreasing stick probabilities encourage each region to activate only a finite subset of the truncated library.

Since b_{ik} is unobserved, we adopt a truncated mean-field posterior $q(b_{ik}) = \mathrm{Bernoulli}(\phi_{ik})$, where ϕ_{ik} denotes the posterior activation probability. The continuous response s_{ik} is assigned a Gaussian posterior with mean μ_{ik}. Let $\boldsymbol{\Phi} = [\phi_{ik}]$ and $\mathbf{M}_S = [\mu_{ik}]$. Rather than sampling binary activations, we construct the induced representation from the posterior expectation:

$$\mathbf{Z}^{\mathrm{ibp}} = (\boldsymbol{\Phi} \odot \mathbf{M}_S)\mathbf{A}, \qquad \mathbf{Z}^{\mathrm{ind}} = \mathrm{Norm}\left[(1-\eta)\mathbf{Z} + \eta \mathbf{Z}^{\mathrm{ibp}}\right]. \tag{4}$$

The residual connection preserves the distilled semantic anchor while injecting sparse region-adaptive factor responses. The variational posterior also provides a region-level uncertainty and reliability estimate:

$$u_i = \frac{1}{K_{\max} \log 2} \sum\nolimits_{k=1}^{K_{\max}} H(\phi_{ik}), \qquad r_i = \exp(-\beta_{\mathrm{rec}} \widetilde{e}_i^{\mathrm{rec}})(1 - u_i), \tag{5}$$

where $\widetilde{e}_i^{\mathrm{rec}}$ is the normalized reconstruction error. A high r_i indicates that the selected factors confidently explain the observed region feature. Region-level reliability is subsequently aggregated within each fragment and used by GM-OT.

Teacher-quality causal calibration: Although the IBP posterior discovers active factors from the unlabeled feature distribution, statistical activation does not necessarily indicate their contribution to pseudo-label construction. We therefore introduce teacher-quality causal calibration to evaluate and adjust their task-level contributions. Specifically, given a fixed probe assignment $\widehat{\mathbf{Y}}$, we suppress the posterior activation of each selected factor k and measure the resulting change in a label-free teacher-quality proxy:

$$\begin{aligned} Q(\mathbf{Z}, \widehat{\mathbf{Y}}) &= M(\mathbf{Z}, \widehat{\mathbf{Y}}) + \lambda_b H_{\mathrm{slot}}(\widehat{\mathbf{Y}}) - \lambda_p C_{\mathrm{proto}}(\mathbf{Z}, \widehat{\mathbf{Y}}), \\ \phi_{ik}^{(-k)} &= \delta \phi_{ik}, \qquad e_k = Q(\mathbf{Z}^{\mathrm{ind}}, \widehat{\mathbf{Y}}) - Q(\mathbf{Z}^{(-k)}, \widehat{\mathbf{Y}}). \end{aligned} \tag{6}$$

Here, M measures the average top-two assignment margin, H_{slot} measures the diversity of primitive-slot usage, and C_{proto} measures prototype correlation and redundancy. The probe assignment is kept

unchanged during intervention, so e_k reflects the representation change caused by suppressing factor k. A positive e_k indicates that factor k contributes positively to pseudo-label construction, whereas a smaller or negative effect indicates a limited contribution. We inject the normalized effect into the posterior logits as $\operatorname{logit} \widetilde{\phi}_{ik} = \operatorname{logit} \phi_{ik} + \lambda_c \bar{e}_k$. The calibrated posterior strengthens factors with larger positive effects and reduces the influence of factors with limited contributions. To prevent adverse interactions among jointly calibrated factors, the resulting representation is accepted only when its overall teacher-quality score does not decrease; otherwise, $\mathbf{Z}^{\mathrm{ind}}$ is retained. The accepted representation is denoted by $\mathbf{Z}^{\mathrm{final}}$.

Complete factor construction, response-aligned ordering, variational updates, and intervention details are provided in the supplementary material.

Spectral Candidate Fragment Discovery: Using $\mathbf{Z}^{\mathrm{final}}$, we construct a spectral embedding (Ng et al., 2001) and overcluster the neural regions into M candidate fragments, where $M > K$. The resulting set $\mathcal{F} = \{\mathcal{F}_j\}_{j=1}^{M}$ preserves fine-grained structural candidates before their assignment to K primitive slots. Each fragment is represented by its induced-feature prototype $\mathbf{o}_j$ and the mean posterior reliability r_j^F of its member regions, which are passed to GM-OT.

3.3 Global Margin-aware Optimal Transport

The candidate fragments in $\mathcal{F}$ must be consolidated into K unordered primitive slots. A local similarity-based rule determines each fragment assignment independently and therefore ignores global competition among fragments and the capacity of different primitive slots. Moreover, the primitive prototypes are estimated from unlabeled fragments rather than semantic annotations. An initially dominant slot may therefore absorb feature-similar ambiguous fragments and progressively reinforce its prototype bias. Fragments that could otherwise support weak structures are consequently diverted to dominant slots, while insufficiently supported slots may be occupied by additional fragments from similar dominant structures, resulting in redundant primitive partitions.

We formulate this merge-back process as a globally constrained transport problem (Peyré & Cuturi, 2019). GM-OT jointly optimizes all fragment-to-primitive assignments under fragment-mass and primitive-capacity constraints, penalizes low-margin and structurally risky assignments, and re-estimates the primitive prototypes from the resulting transport plan.

Transport Marginals and Slot Capacities: GM-OT represents the fragment-to-slot assignment by a non-negative transport plan $\mathbf{P} \in \mathbb{R}_+^{M \times K}$, where P_{jk} is the mass transported from candidate fragment $\mathcal{F}_j$ to primitive slot k. Its row and column sums define the source and target marginals: $\sum_{k=1}^{K} P_{jk} = a_j$ and $\sum_{j=1}^{M} P_{jk} = b_k$, respectively. Thus, a_j specifies the total mass sent by fragment $\mathcal{F}_j$, whereas b_k constrains the total mass received by primitive slot k.

For each fragment, we compute an unnormalized transport-mass score ω_j from its physical support, i.e., the number of represented points, and its within-fragment feature compactness. We further apply risk-aware and frequency corrections to preserve weak or ambiguous fragments and prevent a large group of similar fragments from dominating the source distribution. Here, structural risk summarizes weak physical support, local assignment ambiguity, and the tendency of a fragment to be attracted to a provisionally dominant slot. A provisional grouping provides an initial slot-mass distribution $\widehat{\mathbf{b}}$. The transport marginals are then defined as

$$a_j = \frac{\omega_j}{\sum_{\ell=1}^{M} \omega_\ell}, \qquad b_k = \beta \widehat{b}_k + (1-\beta)\frac{1}{K}, \tag{7}$$

where $j = 1, \ldots, M$, $k = 1, \ldots, K$, and $\beta \in [0, 1]$ controls the contribution of the provisional slot mass. The uniform component prevents a few initially dominant slots from monopolizing the transported mass. The provisional grouping initializes only the slot prototypes and capacities, whereas the final fragment-to-slot assignments are jointly determined by GM-OT.

Fragment-to-Slot Transport Cost: The transport marginals constrain how much mass each fragment sends and each primitive slot receives, but the destination of each fragment is determined by the pairwise transport cost. Given the current prototype $\mathbf{p}_k$ of primitive slot k, the induced-feature distance and competitive assignment margin are defined as

$$d_{jk} = 1 - \cos(\mathbf{o}_j, \mathbf{p}_k), \qquad m_{jk} = \min_{k' \neq k} d_{jk'} - d_{jk}. \tag{8}$$

Here, m_{jk} measures the distance advantage of slot k over its strongest competitor for fragment $\mathcal{F}_j$. A large positive margin indicates a well-separated assignment, whereas a small or negative margin indicates strong competition between primitive slots.

Assignments whose margins fall below the target margin ρ are penalized by

$$C_{jk}^{\mathrm{mar}} = \gamma_j \tau \,\mathrm{softplus}\left(\frac{\rho - m_{jk}}{\tau}\right), \tag{9}$$

where τ controls the smoothness of the penalty. The modulation factor γ_j adapts the margin penalty according to fragment confidence and structural competition. Reliable fragments with ambiguous or structurally risky assignments receive stronger margin enforcement, whereas highly uncertain fragments contribute less to the margin term.

The normalized structural risk $\xi_j \in [0, 1]$ summarizes weak support, local ambiguity, and provisional dominant-slot pressure. It defines the risk-aware assignment term $C_{jk}^{\mathrm{risk}} = \xi_j \bar{d}_{jk}$, where $\bar{d}_{jk}$ is the normalized induced-feature distance. We further use C_{jk}^{dec} to penalize assignments that remain compatible with multiple competing prototypes, thereby discouraging redundant primitive slots. The complete transport cost is

$$C_{jk} = d_{jk} + \lambda_{\mathrm{mar}} C_{jk}^{\mathrm{mar}} + \lambda_{\mathrm{risk}} C_{jk}^{\mathrm{risk}} + \lambda_{\mathrm{dec}} C_{jk}^{\mathrm{dec}}, \tag{10}$$

where the three coefficients balance the margin, risk, and redundancy terms. Given the resulting cost matrix $\mathbf{C}$ and transport marginals $\mathbf{a}$ and $\mathbf{b}$, we use Sinkhorn iterations (Cuturi, 2013) to obtain an entropy-regularized minimum-cost transport plan while iteratively refining the primitive prototypes through transport-weighted fragment aggregation. Complete definitions of the transport marginals and cost components, together with the Sinkhorn and prototype-refinement procedures, are provided in the supplementary material.

This global solution jointly coordinates all fragment-to-slot assignments under the prescribed slot capacities, rather than assigning each fragment independently. After convergence, the hard assignment $\widehat{y}_j^F = \arg\max_k P_{jk}^*$ is propagated from fragment $\mathcal{F}_j$ to its constituent points to supervise the segmentation network. Both modules operate only during pseudo-label construction.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on ScanNet, S3DIS, and nuScenes. **1) ScanNet**: ScanNet contains 1,201 training, 312 validation, and 100 hidden-test scenes with 20 semantic categories. Since the test annotations are withheld, following the evaluation protocol of LogoSP Zhang et al. (2025b) we report the main results on the validation split (Dai et al., 2017). **2) S3DIS** (Armeni et al., 2016): Following the standard protocol, we report results under both the Area-5 and 6-fold cross-validation settings on 12 semantic categories after removing the clutter class (Zhang et al., 2023; 2025b). **3) nuScenes**: nuScenes (Caesar et al., 2020) provides outdoor LiDAR scans collected in autonomous driving scenarios. We follow the standard training/validation split adopted by existing unsupervised methods (Zhang et al., 2025b).

Baseline. We compare with **1) K-means**; adapted 2D methods, **2) IIC** (Ji et al., 2019) and **PiCIE** (Cho et al., 2021); and 3D methods, **3) GrowSP** (Zhang et al., 2023), **4) PointDC** (Chen et al., 2023), **5) PointDC-DINOv2**, **6) LogoSP** (Zhang et al., 2025b), and **7) PointGS** (Song et al., 2026). PointDC-DINOv2 replaces the original 2D prior with DINOv2 (Oquab et al., 2024). The supervised reference methods also include **DGCNN** (Wang et al., 2019), **PointCNN** (Li et al., 2018), **Cylinder3D** (Zhu et al., 2021), and **SPVNAS** (Tang et al., 2020). The supervised transfer comparison additionally includes the linear-probing results of **Utonia** (Zhang et al., 2026), **Sonata** (Wu et al., 2025), and **Concerto** (Zhang et al., 2025a).

Metrics. We report overall accuracy **(OA)**, mean class accuracy **(mAcc)**, and mean Intersection-over-Union **(mIoU)**. Hungarian matching is applied only during evaluation to align unordered primitive indices with semantic classes (Munkres, 1957).

Implementation Details. Our framework is implemented in PyTorch (Paszke et al., 2019), with the 3D segmentation network built using a sparse convolution library (Choy et al., 2019). We adopt

Table 1: ScanNet **validation split** results.

Method	OA (%)	mAcc (%)	mIoU (%)
K-means	10.1	10.0	3.4
IIC (ICCV'19)	27.7	6.1	2.9
PiCIE (CVPR'21)	20.4	16.5	7.6
GrowSP (CVPR'23)	57.3	44.2	25.4
PointDC (ICCV'23)	63.7	–	25.7
PointDC-DINOv2	64.7	45.0	29.6
LogoSP (CVPR'25)	64.7	50.8	35.8
PointGS (CVPR'26)	–	–	36.7
Ours	**72.5**	**55.5**	**41.3**

Table 2: ScanNet **hidden split** results.

	Method	mIoU (%)
Sup.	PointNet++ (NIPS'17)	33.9
	DGCNN (ACM TOG'19)	44.6
	PointCNN (NIPS'18)	45.8
	SparseConv (CVPR'18)	72.5
Unsup.	GrowSP (CVPR'23)	26.9
	PointDC (ICCV'23)	22.9
	LogoSP (CVPR'25)	32.7
	PointGS (CVPR'26)	–
	Ours	**39.8**

Table 3: Quantitative results (%) on S3DIS under the **Area-5** and **6-fold** cross-validation settings.

	Method	$\text{OA}_{\text{Area 5}}$	$\text{mAcc}_{\text{Area 5}}$	$\text{mIoU}_{\text{Area 5}}$	$\text{OA}_{\text{6-fold}}$	$\text{mAcc}_{\text{6-fold}}$	$\text{mIoU}_{\text{6-fold}}$
Sup.	PointNet (CVPR'17)	77.5	59.1	44.6	75.9	67.1	49.4
	PointNet++ (NIPS'17)	77.5	62.6	50.1	77.1	74.1	55.1
	SparseConv (CVPR'18)	88.4	69.2	60.8	89.4	78.1	69.2
Unsup.	K-means	21.4	21.2	8.7	20.0	21.5	8.8
	IIC (ICCV'19)	28.5	12.5	6.4	32.8	14.7	8.5
	PiCIE (CVPR'21)	61.6	25.8	17.9	46.4	28.1	17.8
	GrowSP (CVPR'23)	78.4	57.2	44.5	76.0	59.4	44.6
	PointDC (ICCV'23)	54.1	–	22.6	55.7	37.7	26.0
	PointDC-DINOv2	75.7	48.7	40.2	74.4	51.5	41.3
	LogoSP (CVPR'25)	**82.8**	55.9	46.5	**79.2**	58.0	46.3
	PointGS (CVPR'26)	76.6	66.1	49.3	–	–	45.8
	Ours	81.5	**66.6**	**52.8**	78.9	**69.5**	**51.7**

Res16FPN18 as the 3D backbone and initialize it with DINO/DINOv2-based (Caron et al., 2021; Oquab et al., 2024) 2D-to-3D distillation weights (Chen et al., 2023; Zhang et al., 2025b) to obtain point-level features with semantic priors. We set the residual induction coefficient to $\eta = 0.1$ and the GM-OT cost weights to $(\lambda_{\text{mar}}, \lambda_{\text{risk}}, \lambda_{\text{dec}}) = (0.09, 0.10, 0.15)$. Their sensitivity is analyzed in Appendix.

4.2 Experimental Results

ScanNet. We evaluate on the offline validation and online hidden test sets, excluding undefined points from training and evaluation. Our method achieves the best performance among the compared unsupervised approaches on both splits (Tables 1 and 2), supporting its effectiveness across evaluation protocols.

S3DIS. Our method achieves the highest mIoU and mAcc among the compared unsupervised approaches under both Area-5 and 6-fold evaluation, surpassing supervised PointNet (Table 3). The 6-fold results support generalization across indoor layouts. Qualitative results on ScanNet and S3DIS (Fig. 3) show closer agreement with ground truth in the highlighted regions.

nuScenes. We evaluate on nuScenes (Caesar et al., 2020), which features long-range sparsity, dynamic objects, and severe class imbalance. Our method achieves 22.7% mIoU and 37.3% mAcc, improving on the strongest previous unsupervised results by 2.6 and 7.9 percentage points, respectively, with comparable OA (Table 4). The mAcc gain suggests improved average class recognition under sparse and imbalanced outdoor conditions. Class-wise results for all three benchmarks and additional region-wise S3DIS results are provided in the supplementary material.

4.3 Transferability to Supervised Segmentation

To evaluate the transferability of BN-LFI beyond unsupervised pseudo-label construction, we integrate it into Utonia and conduct supervised evaluation with a frozen encoder. Specifically, Utonia's pretrained encoder and original prediction parameters remain fixed. BN-LFI induces task-relevant active latent factors from its point-level features and calibrates them using supervision from the training set instead of the unsupervised teacher-quality signal. The calibrated representations then refine downstream predictions through a lightweight adaptation module. This process uses only the training set and does not access validation labels.

As shown in Table 5, integrating BN-LFI improves Utonia's mIoU from 77.7 to 78.2 on ScanNet and from 74.7 to 74.9 on S3DIS Area 5. The corresponding mAcc scores increase to 87.4 and

Table 4: Quantitative results on the nuScenes validation split.

	Method	OA	mAcc	mIoU
Sup.	MinkUNet (CVPR'19)	–	–	73.3
	Cylinder3D (CVPR'21)	–	–	76.1
	SPVNAS (ECCV'20)	–	–	77.4
Unsup.	GrowSP (CVPR'23)	39.2	17.5	10.2
	PointDC (ICCV'23)	**56.8**	29.4	17.7
	PointDC-DINOv2	51.8	28.6	18.2
	LogoSP (CVPR'25)	54.8	29.2	20.1
	Ours	54.7	**37.3**	**22.7**

Table 5: Supervised transfer performance (lin.) on ScanNet Val and S3DIS Area 5.

	ScanNet Val			S3DIS Area 5		
Method	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
Sonata (CVPR'25)	72.5	83.1	89.7	72.3	81.2	90.9
Concerto (NIPS'25)	77.3	86.6	91.7	73.5	81.3	90.9
Utonia (ICML'26)	77.7	87.0	91.9	74.7	81.2	**91.5**
Ours	**78.2**	**87.4**	**92.1**	**74.9**	**81.7**	91.4

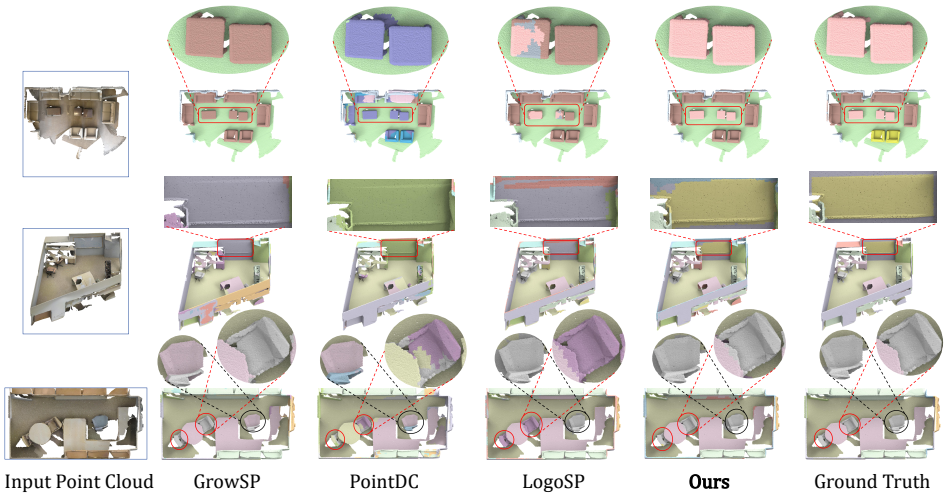


Figure 3: Qualitative results on ScanNet and S3DIS. Colors denote semantic classes; highlighted regions show differences between predictions and ground truth.

81.7, respectively. Meanwhile, allAcc rises from 91.9 to 92.1 on ScanNet. These results support the transferability of BN-LFI to supervised segmentation, where its factor induction and calibration enable lightweight adaptation while keeping Utonia's pretrained encoder and original prediction parameters fixed.

4.4 Ablation Study

Table 6: Ablation study

Factor Induction	Causal Calibration	Fragment Allocation	Margin-aware OT	mIoU(%)
-	-	-	-	32.65
✓	-	-	-	36.01
✓	✓	-	-	37.24
✓	-	✓	-	37.38
✓	✓	✓	-	38.46
✓	✓	✓	✓	**41.33**

We progressively evaluate the components of BN-LFI and GM-OT on the ScanNet validation set. As shown in Table 6, factor induction improves the baseline from 32.65% to 36.01% mIoU, while causal calibration further increases it to 37.24%, validating both data-driven factor discovery and task-level calibration. Introducing candidate-fragment allocation improves factor induction from 36.01% to 37.38%, and its combination with causal calibration reaches 38.46%. Finally, GM-OT provides a further 2.87-point gain and achieves 41.33%, demonstrating that globally constrained assignment complements the induced representation.

5 Conclusion

We presented IFOT-SSeg, a latent factor-driven framework for unsupervised 3D semantic segmentation. BN-LFI induces sparse region-adaptive factors and calibrates their contribution to pseudo-label construction, while GM-OT globally assigns candidate fragments under capacity and competitive-margin constraints. Experiments on ScanNet, S3DIS, and nuScenes demonstrate that the two components produce more reliable pseudo-labels without introducing additional inference overhead.

AI USE STATEMENT

Generative AI tools, including ChatGPT and Google Gemini, were used to assist with language editing, manuscript organization, LaTeX and table formatting, and limited code debugging, including syntax checking, error diagnosis, and troubleshooting of implementation issues. These tools also provided limited feedback on the clarity, consistency, and exposition of the methodological and experimental descriptions. They were not used to generate synthetic data, formulate mathematical claims or proofs, propose research hypotheses, design the core methodology or experimental protocol, independently implement the proposed method, analyze experimental data, or determine the interpretation and conclusions of the results. All research ideas, mathematical formulations, algorithmic designs, core implementations, experiments, and empirical conclusions were developed and validated by the authors. All AI-assisted text, code-related suggestions, and visual content were manually reviewed, revised, tested, and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work develops an unsupervised method for 3D point cloud semantic segmentation and does not involve the collection of new data, human-subject experiments, or personally identifiable information. All experiments are conducted on publicly available research datasets, including ScanNet, S3DIS, and nuScenes, in accordance with their respective licenses and terms of use. The proposed method is intended to advance general-purpose 3D scene understanding. Nevertheless, 3D perception techniques may be deployed in safety-critical or privacy-sensitive applications, such as autonomous driving, robotics, and environmental monitoring. Their performance may also be affected by biases in the geographic, environmental, and semantic distributions of the training datasets. Therefore, deployment in real-world systems should include application-specific validation, privacy safeguards, bias assessment, and appropriate human oversight. The authors do not intend the proposed method for surveillance, identification, or other harmful uses.

REPRODUCIBILITY STATEMENT

We provide the methodological formulations and implementation details necessary to reproduce the proposed framework in the main paper and the appendix. These include the construction and inference procedures of BN-LFI, the cost formulation and optimization procedure of GM-OT, the network architecture, dataset preprocessing, training and optimization settings, hyperparameter configurations, and evaluation protocols. The appendix further reports additional implementation details and sensitivity analyses. All experiments are conducted on publicly available datasets using their standard data splits and evaluation metrics. The authors will release the implementation and experiment configurations.

REFERENCES

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Zisheng Chen, Hongbin Xu, Weitao Chen, Zhipeng Zhou, Haihong Xiao, Baigui Sun, Xuansong Xie, and Wenxiong Kang. Pointdc: Unsupervised semantic segmentation of 3d point clouds via cross-modal distillation and super-voxel clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14290–14299, 2023.

Ran Cheng, Ryan Razani, Yuan Ren, and Liu Bingbing. S3net: 3d lidar sparse semantic segmentation network. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14040–14046, 2021a. doi: 10.1109/ICRA48506.2021.9561305.

Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12547–12556, June 2021b.

Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16794–16804, 2021.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3075–3084, 2019. doi: 10.1109/CVPR.2019.00319.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

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

Finale Doshi, Kurt Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the indian buffet process. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 137–144, 2009.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9224–9232, 2018.

Thomas L. Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(32):1185–1224, 2011.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9865–9874, 2019.

David Knowles and Zoubin Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Independent Component Analysis and Signal Separation*, volume 4666 of *Lecture Notes in Computer Science*, pp. 381–388. Springer, 2007. doi: 10.1007/978-3-540-74494-8_48.

Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4558–4567, 2018.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. PointCNN: Convolution on $\mathcal{X}$-transformed points. In *Advances in Neural Information Processing Systems*, volume 31, pp. 828–838, 2018.

Jiaxu Liu, Zhengdi Yu, Toby P. Breckon, and Hubert P. H. Shum. U3DS^3: Unsupervised 3D semantic scene segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3759–3768, 2024.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14, pp. 849–856, 2001.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

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

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

Yixiao Song, Qingyong Li, Wen Wang, and Zhicheng Yan. PointGS: Semantic-consistent unsupervised 3D point cloud segmentation with 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33343–33352, 2026.

Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3D architectures with sparse point-voxel convolution. In *Computer Vision – ECCV 2020*, volume 12373 of *Lecture Notes in Computer Science*, pp. 685–702. Springer, 2020. doi: 10.1007/978-3-030-58604-1_41.

Yee Whye Teh, Dilan Grür, and Zoubin Ghahramani. Stick-breaking construction for the indian buffet process. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 556–563, 2007. URL https://proceedings.mlr.press/v2/teh07a.html.

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):146:1–146:12, 2019. doi: 10.1145/3326362.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22193–22204, June 2025.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=fV46JjrrOm.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=f2CGTzHQ4E.

Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17619–17629, 2023.

Zihui Zhang, Weisheng Dai, Hongtao Wen, and Bo Yang. Logosp: Local-global grouping of superpoints for unsupervised semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1374–1384, 2025b.

Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9939–9948, 2021.

APPENDIX CONTENTS

A Additional Details of BN-LFI **15**

A.1 ICA-Style Factor Construction 15

A.2 Response-Aligned Factor Ordering 15

A.3 Truncated Variational Posterior 16

A.4 Posterior Reliability 16

A.5 Teacher-Quality Causal Calibration 17

B Additional Details of GM-OT **18**

B.1 Fragment Statistics, Provisional Slots, and Transport Marginals 18

B.2 Reliability-, Risk-, and Margin-Aware Assignment Costs 19

B.3 Transport Optimization and Discretization 19

C Pseudo-Label Training **20**

D Implementation Details and Hyperparameter Analysis **20**

D.1 Reproducibility Settings 20

D.2 Hyperparameter Sensitivity 21

D.3 Robustness to Training Randomness 21

D.4 Computational Complexity and Runtime 22

E Algorithm Details **22**

F Additional Experimental Results **23**

F.1 Evaluation on ScanNet 23

F.2 Evaluation on S3DIS 24

F.3 Evaluation on nuScenes 28

A Additional Details of BN-LFI

This section details the factor construction, response-aligned ordering, variational inference, posterior reliability, and teacher-quality calibration used by BN-LFI.

A.1 ICA-Style Factor Construction

Following standard whitening and ICA principles (Hyvärinen & Oja, 2000; Knowles & Ghahramani, 2007), we compute the whitened representation of the centered neural-region feature matrix $\bar{\mathbf{Z}}$ as

$$\mathbf{Z}^w = \mathrm{Norm}\left(\bar{\mathbf{Z}}\boldsymbol{\Sigma}^{-1/2}\right), \tag{11}$$

where $\boldsymbol{\Sigma}^{-1/2}$ is approximated using Newton–Schulz iterations. Whitening reduces second-order correlations among the observed channels and provides a stable coordinate space for subsequent factor construction.

Let $\mathbf{X} = \mathrm{Norm}(\mathbf{Z}^w) \in \mathbb{R}^{N\times D}$ denote the row-normalized whitened neural-region features. We initialize $K_{\max}$ unit-norm candidate directions by randomized SVD followed by Varimax rotation. When $K_{\max} > D$, additional directions are generated as normalized random mixtures of the rotated basis. These directions initialize the overcomplete library $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{K_{\max}}]^\top$.

The candidate library is refined using alternating sparse coding and dictionary updates. At iteration t, the response matrix is obtained by soft thresholding,

$$\mathbf{C}^{(t)} = \mathcal{S}_{\lambda_A}\left(\mathbf{X}\mathbf{A}^{(t)\top}\right), \qquad \mathcal{S}_\lambda(x) = \mathrm{sign}(x)\max(|x| - \lambda, 0). \tag{12}$$

For factor k, its data-supported update is

$$\widehat{\mathbf{a}}_k^{(t+1)} = \mathrm{Norm}\left(\frac{\sum_{i=1}^N C_{ik}^{(t)}\mathbf{x}_i}{\sum_{i=1}^N |C_{ik}^{(t)}| + \epsilon_A}\right). \tag{13}$$

We additionally subtract the response correlation contributed by the other factors and blend the update with the previous direction,

$$\mathbf{A}^{(t+1)} = \mathrm{Norm}\Big[(1-\xi_A)\mathbf{A}^{(t)} + \xi_A\,\mathrm{Norm}\Big(\widehat{\mathbf{A}}^{(t+1)} - \lambda_A^{\mathrm{dec}}\mathbf{G}_C^{(t)}\mathbf{A}^{(t)}\Big)\Big], \tag{14}$$

where $\mathbf{G}_C^{(t)}$ is the off-diagonal correlation matrix of the standardized columns of $\mathbf{C}^{(t)}$. Factors with negligible response mass are restored to their initial directions to avoid unstable updates. This procedure yields sparse responses while discouraging redundant candidate directions; it does not use semantic annotations.

A.2 Response-Aligned Factor Ordering

ICA-style factors are permutation-indeterminate, whereas the stick-breaking IBP prior is size-biased (Teh et al., 2007; Griffiths & Ghahramani, 2011). To prevent arbitrary factor indices from determining the initial activation prior, we align the factor order with empirical support. For the refined library, we compute

$$R_{ik} = \mathbf{z}_i^{w\top}\mathbf{a}_k, \qquad C_{ik} = \mathcal{S}_{\lambda_A}(R_{ik}), \qquad h_k = \frac{1}{N}\sum_{i=1}^N |C_{ik}|. \tag{15}$$

If σ satisfies $h_{\sigma(1)} \geq \cdots \geq h_{\sigma(K_{\max})}$, the factors are reordered before variational inference as

$$\mathbf{A} \leftarrow [\mathbf{a}_{\sigma(1)}, \ldots, \mathbf{a}_{\sigma(K_{\max})}]^\top. \tag{16}$$

This ordering only aligns the initialization of the size-biased prior with data support. The final factor usage remains region-specific and is inferred through $q(b_{ik}) = \mathrm{Bernoulli}(\phi_{ik})$.

A.3 Truncated Variational Posterior

Following truncated variational inference for the IBP (Doshi et al., 2009), for each neural region i and factor k, the generative model assumes

$$v_k \sim \mathrm{Beta}(\alpha, 1), \quad \pi_k = \prod_{\ell=1}^{k} v_\ell, \quad b_{ik} \sim \mathrm{Bernoulli}(\pi_k), \quad s_{ik} \sim \mathcal{N}(0,1), \tag{17}$$

and

$$\mathbf{z}_i^w \mid \mathbf{b}_i, \mathbf{s}_i, \mathbf{A} \sim \mathcal{N}\left(\sum_{k=1}^{K_{\max}} b_{ik} s_{ik} \mathbf{a}_k, \sigma_z^2 \mathbf{I} \right). \tag{18}$$

We use the mean-field family

$$q(\mathbf{V}, \mathbf{B}, \mathbf{S}) = \prod_{k=1}^{K_{\max}} q(v_k) \prod_{i=1}^{N} \prod_{k=1}^{K_{\max}} q(b_{ik}) q(s_{ik}), \tag{19}$$

where $q(v_k) = \mathrm{Beta}(\widetilde{a}_k, \widetilde{b}_k)$, $q(b_{ik}) = \mathrm{Bernoulli}(\phi_{ik})$, and $q(s_{ik})$ is Gaussian. The corresponding variational objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{VI}} = \mathbb{E}_q[\log p(\mathbf{Z}^w \mid \mathbf{B}, \mathbf{S}, \mathbf{A})] - \mathrm{KL}[q(\mathbf{V}) \| p(\mathbf{V})] \\ - \mathbb{E}_{q(\mathbf{V})} \mathrm{KL}[q(\mathbf{B}) \| p(\mathbf{B} \mid \mathbf{V})] - \mathrm{KL}[q(\mathbf{S}) \| p(\mathbf{S})]. \end{aligned} \tag{20}$$

Let $R_{ik} = \mathbf{z}_i^{w\top} \mathbf{a}_k$. Integrating the Gaussian coefficient gives the following evidence gain for activating factor k:

$$\Delta_{ik} = -\frac{1}{2} \log \frac{1+\sigma_z^2}{\sigma_z^2} + \frac{R_{ik}^2}{2} \left(\frac{1}{\sigma_z^2} - \frac{1}{1+\sigma_z^2} \right). \tag{21}$$

At each fixed-point iteration, the activation posterior is updated as

$$\phi_{ik}^{\mathrm{new}} = \mathrm{sigmoid}\left(\mathbb{E}_q[\log \pi_k] - \mathbb{E}_q[\log(1-\pi_k)] + \frac{\Delta_{ik}}{T} + \lambda_c \bar{e}_k \right), \tag{22}$$

Here, T is the posterior temperature, λ_c controls the causal feedback strength, and $\bar{e}_k$ is the normalized factor effect defined below during causal calibration. The causal term is set to zero during the initial BN-LFI inference. We then apply damping as $\phi_{ik} \leftarrow d\phi_{ik} + (1-d)\phi_{ik}^{\mathrm{new}}$. The stick parameters are then updated from the posterior usage of all factors controlled by each stick:

$$\widetilde{a}_\ell = \alpha + \sum_{i=1}^{N} \sum_{k=\ell}^{K_{\max}} \phi_{ik}, \qquad \widetilde{b}_\ell = 1 + \kappa_{\mathrm{stick}} \sum_{i=1}^{N} \sum_{k=\ell}^{K_{\max}} (1-\phi_{ik}), \tag{23}$$

where κ_{stick} damps the accumulated inactive evidence. The continuous coefficient posterior has

$$\mu_{ik} = \frac{R_{ik}}{1+\sigma_z^2}, \qquad \mathrm{Var}_q(s_{ik}) = \frac{\sigma_z^2}{1+\sigma_z^2}. \tag{24}$$

After convergence, the posterior expectation gives

$$\mathbf{Z}^{\mathrm{ibp}} = (\mathbf{\Phi} \odot \mathbf{M}_S)\mathbf{A}, \qquad \mathbf{Z}^{\mathrm{ind}} = \mathrm{Norm}\big[(1-\eta)\mathbf{Z} + \eta \mathbf{Z}^{\mathrm{ibp}}\big]. \tag{25}$$

A.4 Posterior Reliability

The reliability used by GM-OT is computed directly from the variational posterior, without semantic labels. Let $\mathbf{z}_i^{\mathrm{ibp}}$ be the row-normalized posterior reconstruction in Eq. (25). We first compute the per-region mean-squared reconstruction error

$$e_i^{\mathrm{rec}} = \frac{1}{D} \left\| \mathbf{z}_i^w - \mathbf{z}_i^{\mathrm{ibp}} \right\|_2^2, \qquad \widetilde{e}_i^{\mathrm{rec}} = \frac{e_i^{\mathrm{rec}} - \min_n e_n^{\mathrm{rec}}}{\max_n e_n^{\mathrm{rec}} - \min_n e_n^{\mathrm{rec}} + \epsilon_r}. \tag{26}$$

The posterior uncertainty is the average normalized Bernoulli entropy,

$$u_i = \frac{1}{K_{\max} \log 2} \sum_{k=1}^{K_{\max}} H(\phi_{ik}), \qquad H(p) = -p \log p - (1-p)\log(1-p). \tag{27}$$

An activation probability close to zero or one has low uncertainty, whereas a probability close to 0.5 has high uncertainty. We combine reconstruction accuracy and posterior certainty as

$$r_i = \exp(-\beta_{\text{rec}} \widetilde{e}_i^{\text{rec}})(1 - u_i), \tag{28}$$

so a region is reliable only when the selected factors both reconstruct its observation accurately and have a confident activation posterior. For a candidate fragment $\mathcal{F}_j$, the fragment-level reliability and uncertainty used by GM-OT are simple averages over its member regions:

$$r_j^F = \frac{1}{|\mathcal{F}_j|} \sum_{i \in \mathcal{F}_j} r_i, \qquad u_j^F = \frac{1}{|\mathcal{F}_j|} \sum_{i \in \mathcal{F}_j} u_i. \tag{29}$$

A.5 Teacher-Quality Causal Calibration

For computational efficiency, causal calibration intervenes on the K_c factors with the largest global posterior usage among those satisfying a minimum-usage threshold. A probe assignment $\widehat{\mathbf{Y}}$ is kept fixed throughout all interventions. Let $\mathcal{I}_k = \{i : \widehat{y}_i = k\}$ and define the probe prototype

$$\mathbf{p}_k = \text{Norm}\left(\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathbf{z}_i \right). \tag{30}$$

If $s_i^{(1)}$ and $s_i^{(2)}$ are the largest and second-largest values in $\{\cos(\mathbf{z}_i, \mathbf{p}_k)\}_{k=1}^K$, the three components of the label-free teacher quality are

$$\begin{aligned} M(\mathbf{Z}, \widehat{\mathbf{Y}}) &= \frac{1}{N} \sum_{i=1}^{N} \left(s_i^{(1)} - s_i^{(2)} \right), \\ H_{\text{slot}}(\widehat{\mathbf{Y}}) &= -\frac{1}{\log K} \sum_{k:\nu_k > 0} \nu_k \log \nu_k, \qquad \nu_k = \frac{|\mathcal{I}_k|}{N}, \\ C_{\text{proto}}(\mathbf{Z}, \widehat{\mathbf{Y}}) &= \frac{1}{K(K-1)} \sum_{k \neq \ell} \left| \mathbf{p}_k^\top \mathbf{p}_\ell \right|. \end{aligned} \tag{31}$$

The complete quality proxy is

$$Q(\mathbf{Z}, \widehat{\mathbf{Y}}) = M(\mathbf{Z}, \widehat{\mathbf{Y}}) + \lambda_b H_{\text{slot}}(\widehat{\mathbf{Y}}) - \lambda_p C_{\text{proto}}(\mathbf{Z}, \widehat{\mathbf{Y}}). \tag{32}$$

For factor k, we set $\phi_{ik}^{(-k)} = \delta \phi_{ik}$ for all regions and reconstruct the intervened representation $\mathbf{Z}^{(-k)}$. Its effect is

$$e_k = Q(\mathbf{Z}^{\text{ind}}, \widehat{\mathbf{Y}}) - Q(\mathbf{Z}^{(-k)}, \widehat{\mathbf{Y}}). \tag{33}$$

For the set $\mathcal{K}_c$ of intervened factors, the normalized feedback is

$$\bar{e}_k = \frac{e_k - \frac{1}{|\mathcal{K}_c|} \sum_{\ell \in \mathcal{K}_c} e_\ell}{\max_{\ell \in \mathcal{K}_c} \left| e_\ell - \frac{1}{|\mathcal{K}_c|} \sum_{q \in \mathcal{K}_c} e_q \right| + \epsilon_e}. \tag{34}$$

The normalized effect $\bar{e}_k$ is inserted into Eq. (22). Positive effects strengthen factors whose suppression degrades teacher quality, whereas negative effects suppress redundant or harmful responses. The resulting calibrated posterior $\widetilde{\mathbf{\Phi}}$ produces

$$\mathbf{Z}^{\text{cal}} = \text{Norm}\left[(1 - \eta)\mathbf{Z} + \eta(\widetilde{\mathbf{\Phi}} \odot \mathbf{M}_S)\mathbf{A} \right]. \tag{35}$$

The final representation is selected by

$$\mathbf{Z}^{\text{final}} = \begin{cases} \mathbf{Z}^{\text{cal}}, & Q(\mathbf{Z}^{\text{cal}}, \widehat{\mathbf{Y}}) \geq Q(\mathbf{Z}^{\text{ind}}, \widehat{\mathbf{Y}}), \\ \mathbf{Z}^{\text{ind}}, & \text{otherwise.} \end{cases} \tag{36}$$

B Additional Details of GM-OT

B.1 Fragment Statistics, Provisional Slots, and Transport Marginals

For fragment $\mathcal{F}_j$, let $n_i = |\mathcal{R}_i|$ be the number of points represented by region i. Its physical support, induced-feature prototype, and mean cosine compactness are

$$\mu_j = \sum_{i \in \mathcal{F}_j} n_i, \qquad \mathbf{o}_j = \mathrm{Norm}\left(\frac{1}{\mu_j} \sum_{i \in \mathcal{F}_j} n_i \mathbf{z}_i^{\mathrm{final}} \right), \qquad \bar{c}_j = \frac{1}{\mu_j} \sum_{i \in \mathcal{F}_j} n_i \cos(\mathbf{z}_i^{\mathrm{final}}, \mathbf{o}_j). \tag{37}$$

To estimate structural competition before transport, we first perform a coarse K-way grouping in the spectral space used for fragment discovery. We then recompute the corresponding reference prototypes $\{\widetilde{\mathbf{p}}_k\}_{k=1}^K$ in the induced-feature space, so that they can be compared with the fragment prototypes $\mathbf{o}_j$. Each fragment is associated with its nearest reference prototype,

$$h_0(j) = \arg\min_k \left[1 - \cos(\mathbf{o}_j, \widetilde{\mathbf{p}}_k)\right]. \tag{38}$$

The corresponding top-two similarity margin and received physical support are

$$\Delta_j = \cos(\mathbf{o}_j, \widetilde{\mathbf{p}}_{h_0(j)}) - \max_{k \neq h_0(j)} \cos(\mathbf{o}_j, \widetilde{\mathbf{p}}_k), \qquad H_k = \sum_{j: h_0(j) = k} \mu_j. \tag{39}$$

These quantities are used only to estimate fragment risk and do not determine the final GM-OT assignment. Here, $\mathcal{R}_{01}(\cdot)$ denotes robust percentile normalization that maps the fifth and ninety-fifth percentiles to 0 and 1, respectively, and clips values outside this range. For fragment $\mathcal{F}_j$, weak-support risk, assignment ambiguity, and dominant-slot pressure are defined as

$$\begin{aligned} s_j^{\mathrm{size}} &= 1 - \mathcal{R}_{01}(\log(1 + \mu_j)), \\ s_j^{\mathrm{amb}} &= 1 - \mathcal{R}_{01}(\Delta_j), \qquad s_j^{\mathrm{head}} = \mathcal{R}_{01}\big(H_{h_0(j)}\big), \end{aligned} \tag{40}$$

where μ_j is the physical support of the fragment, Δ_j is its top-two similarity margin under the provisional assignment, and $H_{h_0(j)}$ is the total support received by its provisional slot. The structural risk is

$$\xi_j = \mathrm{clip}_{[0,1]}\left(0.45 s_j^{\mathrm{size}} + 0.35 s_j^{\mathrm{head}} + 0.20 s_j^{\mathrm{amb}}\right). \tag{41}$$

We then define

$$c_j = \mathcal{R}_{01}(\bar{c}_j), \qquad s_j = \sqrt{\mathrm{clip}_{[0,1]}(\mu_j / \mu_{.75})}, \qquad q_j = \sqrt{c_j s_j}, \tag{42}$$

where $\mu_{.75}$ is the seventy-fifth percentile of the positive fragment masses. To estimate the prevalence of repeated structural patterns, we partition the fragment prototypes $\{\mathbf{o}_j\}_{j=1}^M$ into R provisional similarity families and denote the family index of fragment j by $f(j)$. The total mass of family r is $M_r = \sum_{j: f(j) = r} \mu_j$. Its frequency pressure is

$$g_j = \max\left[\left(\frac{M_{f(j)}}{\mathrm{median}_{r: M_r > 0} M_r} \right)^{\gamma_f}, 0.35 \right]. \tag{43}$$

The source marginal is constructed as

$$\omega_j = \mathrm{clip}_{[Q_{.02}, Q_{.98}]}\left[\frac{\max(\mu_j, 1)^{\kappa_{\mathrm{mass}}} (0.35 + 0.65 q_j)(1 + \lambda_s \xi_j)}{g_j} \right], \qquad a_j = \frac{\omega_j}{\sum_{\ell=1}^M \omega_\ell}. \tag{44}$$

Here, $Q_{.02}$ and $Q_{.98}$ are the second and ninety-eighth percentiles of the untrimmed scores, κ_{mass} controls the contribution of physical support, λ_s controls structural-risk compensation, and γ_f determines the frequency correction applied to repeatedly observed fragment families. These corrections affect only the transported source mass and do not determine the final fragment-to-slot assignment.

Using the resulting structural masses, we obtain provisional primitive supports through a weighted coarse grouping of the fragment prototypes:

$$\{h(j)\}_{j=1}^M, \{\mathbf{p}_k^{(0)}\}_{k=1}^K = \arg\min_{\{h(j)\}, \{\mathbf{p}_k\}} \sum_{j=1}^M \omega_j \left[1 - \cos(\mathbf{o}_j, \mathbf{p}_{h(j)})\right]. \tag{45}$$

Here, $h(j)$ is the provisional slot of fragment j, and $\mathbf{p}_k^{(0)}$ is the initial prototype of slot k. This coarse grouping only initializes the target capacities and primitive prototypes; it does not determine the final fragment assignment. The normalized initial slot mass is

$$\widehat{b}_k = \sum_{j:h(j)=k} a_j. \tag{46}$$

The target marginal blends this empirical capacity prior with a uniform prior:

$$\widetilde{b}_k = \beta \widehat{b}_k + (1-\beta)\frac{1}{K}, \qquad b_k = \frac{\max(\widetilde{b}_k, b_{\min}/K)}{\sum_{r=1}^{K} \max(\widetilde{b}_r, b_{\min}/K)}. \tag{47}$$

Here, $\beta \in [0,1]$ controls the contribution of the provisional slot mass, and $b_{\min}$ prevents any slot from receiving a vanishing target capacity. The provisional grouping only initializes the slot capacities and prototypes; the final assignments are determined by the global transport plan.

B.2 Reliability-, Risk-, and Margin-Aware Assignment Costs

Let $\mathbf{o}_j$ denote the normalized prototype of fragment $\mathcal{F}_j$, and let $\mathbf{p}_k$ denote the current prototype of primitive slot k, initialized by $\mathbf{p}_k^{(0)}$ and subsequently refined by GM-OT. The induced-feature distance and signed competitive margin are

$$d_{jk} = 1 - \cos(\mathbf{o}_j, \mathbf{p}_k), \qquad m_{jk} = \min_{k' \neq k} d_{jk'} - d_{jk}. \tag{48}$$

The modulation applied to the margin penalty is

$$\gamma_j = (0.25 + 0.75 r_j^F)\left(0.5 + 0.5\max[s_j^{\mathrm{amb}}, \xi_j]\right)(1 - 0.5 u_j^F), \tag{49}$$

where r_j^F and u_j^F are the fragment-level posterior reliability and uncertainty, s_j^{amb} is local assignment ambiguity, and ξ_j is structural risk. The margin cost is

$$C_{jk}^{\mathrm{mar}} = \gamma_j \tau \,\mathrm{softplus}\left(\frac{\rho - m_{jk}}{\tau}\right). \tag{50}$$

The normalized feature distance is $\bar{d}_{jk} = \mathcal{R}_{01}(d_{jk})$, and the structural-risk cost is

$$C_{jk}^{\mathrm{risk}} = \xi_j \bar{d}_{jk}. \tag{51}$$

The redundancy cost penalizes a fragment that remains compatible with several competing prototypes:

$$C_{jk}^{\mathrm{dec}} = \mathcal{R}_{01}\left(\frac{1}{K-1}\sum_{k' \neq k}[\max(0, \mathbf{o}_j^\top \mathbf{p}_{k'})]^2\right). \tag{52}$$

The complete pairwise transport cost is therefore

$$C_{jk} = d_{jk} + \lambda_{\mathrm{mar}} C_{jk}^{\mathrm{mar}} + \lambda_{\mathrm{risk}} C_{jk}^{\mathrm{risk}} + \lambda_{\mathrm{dec}} C_{jk}^{\mathrm{dec}}. \tag{53}$$

B.3 Transport Optimization and Discretization

Following entropy-regularized optimal transport (Cuturi, 2013; Peyré & Cuturi, 2019), we define the feasible transport set for source marginal $\mathbf{a}$ and target marginal $\mathbf{b}$ as

$$\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \geq 0 : \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}\}. \tag{54}$$

GM-OT solves the entropy-regularized transport problem

$$\mathbf{P}^* = \arg\min_{\mathbf{P} \in \Pi(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}\rangle + \varepsilon \sum_{j,k} P_{jk}(\log P_{jk} - 1). \tag{55}$$

We solve Eq. (55) by Sinkhorn scaling. Let $\mathbf{O} \in \mathbb{R}^{M \times D}$ and $\mathbf{U} \in \mathbb{R}^{K \times D}$ collect the fragment and slot prototypes, respectively. After each transport step, the transport-weighted slot prototypes are computed as

$$\widehat{\mathbf{U}} = \mathrm{Norm}_{\mathrm{row}}\left[\mathrm{diag}(\mathbf{P}^{*\top}\mathbf{1})^{-1}\mathbf{P}^{*\top}\mathbf{O}\right]. \tag{56}$$

We then explicitly decorrelate the updated slot prototypes:

$$\widehat{\mathbf{U}}^{\mathrm{dec}} = \mathrm{Norm}_{\mathrm{row}}\left[\widehat{\mathbf{U}} - 4\alpha_{\mathrm{dec}}\left(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top} - \mathbf{I}\right)\widehat{\mathbf{U}}\right]. \tag{57}$$

The refined prototypes are blended with the previous prototypes:

$$\mathbf{U} \leftarrow \mathrm{Norm}_{\mathrm{row}}\left[(1-\zeta)\mathbf{U} + \zeta\widehat{\mathbf{U}}^{\mathrm{dec}}\right]. \tag{58}$$

The cost matrix and transport plan are recomputed after each prototype update.

After the final prototype refinement and transport update, fragment $\mathcal{F}_j$ receives the hard slot assignment

$$\widehat{y}_j^F = \arg\max_k P_{jk}^*. \tag{59}$$

C Pseudo-Label Training

The fragment assignment $\widehat{y}_j^F$ is propagated to all member neural regions and their constituent points. Let $\widehat{y}_i$ denote the resulting pseudo-label of point i, and let Ω contain points with valid pseudo-labels. A standard 3D segmentation network with a cosine classifier is trained using

$$\mathcal{L}_{\mathrm{seg}} = -\frac{1}{|\Omega|}\sum_{i\in\Omega}\log p_i(\widehat{y}_i). \tag{60}$$

Both point features and classifier prototypes are ℓ_2-normalized before the logits are computed. Each classifier prototype is initialized as the mean backbone feature of the points assigned to its primitive slot. During this stage, only the segmentation backbone and classifier are optimized; BN-LFI and GM-OT remain fixed and are not used at inference time.

D Implementation Details and Hyperparameter Analysis

D.1 Reproducibility Settings

We use a shared configuration for the principal components of our method across all datasets. Specifically, the IBP concentration, residual induction coefficient, and target assignment margin are fixed as $\alpha = 4$, $\eta = 0.1$, and $\rho = 0.08$, respectively. The GM-OT cost weights are set to $(\lambda_{\mathrm{mar}}, \lambda_{\mathrm{risk}}, \lambda_{\mathrm{dec}}) = (0.09, 0.10, 0.15)$. The remaining dataset-specific settings are summarized in Table 7.

Table 7: Dataset-specific implementation settings.

Dataset	D	$K_{\max}$	K	Voxel (m)	Batch size
ScanNet	384	512	20	0.05	8
S3DIS	384	512	12	0.05	4
nuScenes	128	256	16	0.15	8

Following the corresponding distilled baselines, we adopt their initial superpoint preprocessing and group feature-similar superpoints into at most 40 neural regions per indoor scene and 80 per nuScenes scan. For all datasets, we overcluster the neural regions into $M = 100$ candidate fragments, while the number of primitive slots K follows the semantic category count of each evaluation protocol, as reported in Table 7.

The segmentation network is trained for 200 epochs using Adam with an initial learning rate of 10^{-4} and polynomial decay with power 0.9. Points with undefined labels are excluded from training and evaluation. Hungarian matching is used exclusively for evaluation and does not participate in pseudo-label construction or network optimization.

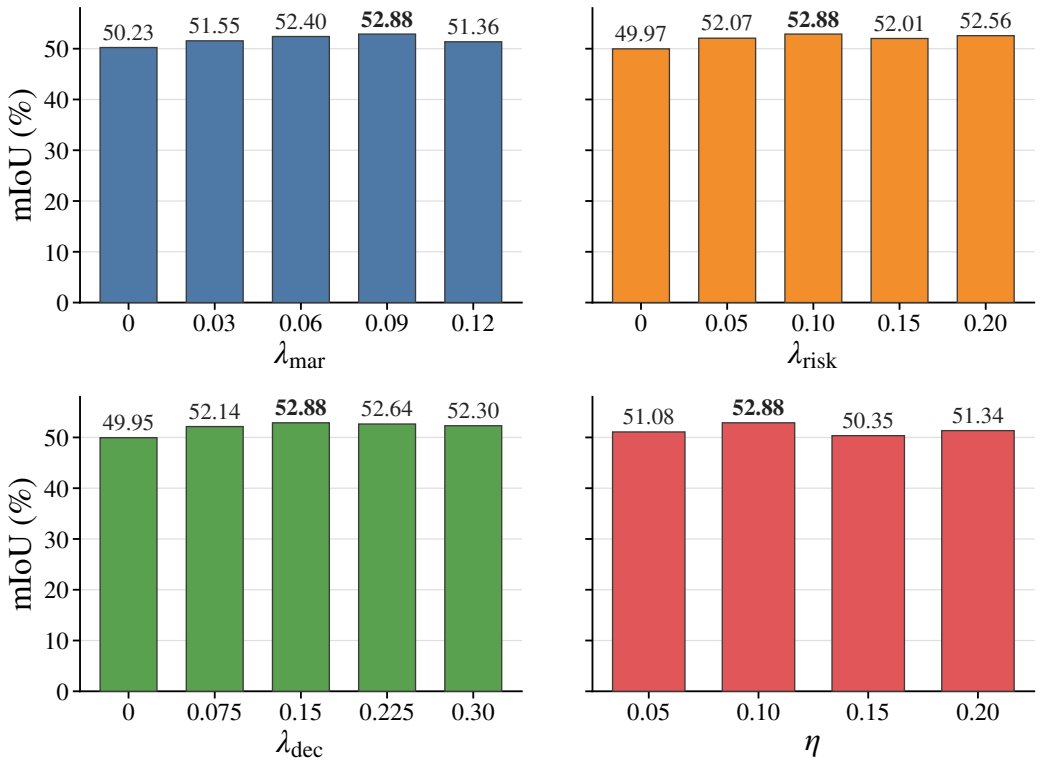


Figure 4: One-at-a-time sensitivity analysis of the residual induction coefficient and the three GM-OT cost weights on S3DIS Area-5. Each parameter is varied while all remaining settings are fixed to the main configuration. Bold values indicate the settings used in the main experiments.

D.2 Hyperparameter Sensitivity

We conduct a one-at-a-time sensitivity analysis on S3DIS Area-5. The residual induction coefficient η and each GM-OT cost weight are varied independently, while all remaining settings are fixed to the main configuration.

GM-OT cost weights. As shown in Fig. 4, disabling any individual cost term consistently degrades performance. Setting λ_{mar}, λ_{risk}, or λ_{dec} to zero reduces mIoU from 52.88% to 50.23%, 49.97%, and 49.95%, respectively, confirming that competitive-margin enforcement, structural-risk modeling, and prototype-redundancy suppression provide complementary contributions. Performance improves as λ_{mar} increases to 0.09 but decreases at 0.12, indicating that excessive margin regularization can override the feature-matching cost. The risk and decorrelation terms remain comparatively stable over a range of nonzero weights, attaining their best results at $\lambda_{\text{risk}} = 0.10$ and $\lambda_{\text{dec}} = 0.15$.

Residual induction coefficient. The coefficient η controls the contribution of the factor-induced residual relative to the distilled feature anchor. A small value ($\eta = 0.05$) does not fully exploit the induced factors, whereas larger values place excessive weight on the reconstructed representation and reduce performance. The intermediate setting $\eta = 0.10$ obtains the best result of 52.88% mIoU. We therefore use $(\lambda_{\text{mar}}, \lambda_{\text{risk}}, \lambda_{\text{dec}}, \eta) = (0.09, 0.10, 0.15, 0.10)$ in the main experiments.

D.3 Robustness to Training Randomness

We evaluate training robustness on S3DIS Area-5 using five random seeds. As shown in Table 8, IFOT-SSeg obtains $(52.42 \pm 0.32)\%$ mIoU, demonstrating stable performance across runs.

Table 8: Robustness to random seeds on S3DIS Area-5. Mean and sample standard deviation are reported over five runs.

Seed	OA (%)	mAcc (%)	mIoU (%)
2026	82.08	62.30	52.68
2027	81.01	63.45	52.29
2028	81.80	62.16	52.81
2029	80.77	62.28	52.00
2030	82.05	61.19	52.32
Mean $\pm$ std.	81.54 ± 0.61	62.27 ± 0.80	52.42 ± 0.32

D.4 Computational Complexity and Runtime

We analyze the one-time pseudo-label construction cost separately from the downstream segmentation network. Denote the number of neural regions, observed feature dimension, truncated factor count, sampled regions for spectral construction, candidate fragments, and primitive slots by N, D, $K_{\max}$, S, M, and K, respectively. We further use T_w, T_A, T_V, T_O, and T_S for the numbers of whitening, factor-library, variational, transport-refinement, and Sinkhorn iterations, and F for the number of intervened factors. Table 9 summarizes the asymptotic costs of the pseudo-label construction stages.

Table 9: Time and space complexity of pseudo-label construction. The spectral candidate stage is inherited from the underlying overclustering pipeline.

Component	Time complexity	Space complexity
Whitening	$\mathcal{O}(ND^2 + T_w D^3)$	$\mathcal{O}(ND + D^2)$
Factor-library construction	$\mathcal{O}\big(T_A(NDK_{\max} + NK_{\max}^2 + K_{\max}^2 D)\big)$	$\mathcal{O}(NK_{\max} + K_{\max}D + K_{\max}^2)$
Variational IBP inference	$\mathcal{O}(NDK_{\max} + T_V NK_{\max})$	$\mathcal{O}(NK_{\max} + K_{\max}D)$
Causal calibration	$\mathcal{O}(FNDK_{\max})$	$\mathcal{O}(NK_{\max} + K_{\max}D)$
Spectral candidate construction	$\mathcal{O}(S^2D + S^3)$	$\mathcal{O}(S^2 + SD)$
GM-OT	$\mathcal{O}(T_O(MKD + T_S MK))$	$\mathcal{O}(MK + (M+K)D)$

Table 10 reports wall-clock measurements from the S3DIS Area-5 experiment on one NVIDIA GeForce RTX 4080 SUPER GPU. With $(N, D, K_{\max}, S, M, K) = (7{,}989, 384, 512, 5{,}000, 100, 12)$, complete pseudo-label construction takes 239.67 seconds, approximately 12.5% of the time required for 200-epoch segmentation training. In particular, GM-OT operates on only a 100×12 transport plan, so its additional memory is small relative to factor inference and dense spectral construction. BN-LFI and GM-OT are removed after pseudo-label generation; consequently, they add no parameters, memory, or computation to the deployed segmentation model.

Table 10: Wall-clock cost on S3DIS Area-5. Pseudo-label construction is performed once before segmentation training.

Stage	Time (s)
Region grouping and BN-LFI	17.00
Teacher-quality causal calibration	2.93
Candidate fragments and GM-OT	219.74
Complete pseudo-label construction	239.67
Segmentation training (200 epochs)	1,923.28

E Algorithm Details

Algorithm 1 summarizes the complete pseudo-teacher construction and downstream segmentation procedure. BN-LFI and GM-OT are executed only during pseudo-label construction; the final segmentation network is trained with the resulting hard pseudo-labels and does not invoke either module at inference time.

Algorithm 1 IFOT-SSeg Pseudo-label Construction and Segmentation Training

Require: Unlabeled point clouds $\mathcal{P}$; distilled encoder E_ψ; primitive number K; overclustering number M; factor truncation $K_{\max}$
Ensure: Point-level pseudo-labels $\widehat{\mathbf{y}}$ and trained segmentation network S_θ

Extract point-wise features and group them into neural regions: $\mathbf{Z} \leftarrow \mathcal{R}(E_\psi(\mathcal{P}))$
▷ **BN-LFI: latent factor induction**
Whiten $\mathbf{Z}$ and construct an overcomplete candidate factor library $\mathbf{A} = \{\mathbf{a}_k\}_{k=1}^{K_{\max}}$
Order candidate factors according to their data-supported response mass
Initialize the truncated variational IBP posterior
repeat
    Update region-factor activation probabilities $\mathbf{\Phi}$
    Update continuous factor responses $\mathbf{M}_S$ and stick posterior
until the variational posterior converges
Construct the induced representation:

$$\mathbf{Z}^{\mathrm{ind}} \leftarrow \mathrm{Norm}\left[(1-\eta)\mathbf{Z} + \eta(\mathbf{\Phi} \odot \mathbf{M}_S)\mathbf{A}\right]$$

Compute region reliability $\{r_i\}$ and posterior uncertainty $\{u_i\}$
▷ **Teacher-quality causal calibration**
Construct a fixed probe assignment $\widehat{\mathbf{Y}}$ from $\mathbf{Z}^{\mathrm{ind}}$
for each selected active factor k **do**
    Suppress factor k and obtain the intervened representation $\mathbf{Z}^{(-k)}$
    Evaluate its contribution:

$$e_k \leftarrow Q(\mathbf{Z}^{\mathrm{ind}}, \widehat{\mathbf{Y}}) - Q(\mathbf{Z}^{(-k)}, \widehat{\mathbf{Y}})$$

end for
Calibrate the activation logits using the normalized factor effects
Reconstruct the calibrated representation $\mathbf{Z}^{\mathrm{cal}}$
if $Q(\mathbf{Z}^{\mathrm{cal}}, \widehat{\mathbf{Y}}) \geq Q(\mathbf{Z}^{\mathrm{ind}}, \widehat{\mathbf{Y}})$ **then**
    $\mathbf{Z}^{\mathrm{final}} \leftarrow \mathbf{Z}^{\mathrm{cal}}$
else
    $\mathbf{Z}^{\mathrm{final}} \leftarrow \mathbf{Z}^{\mathrm{ind}}$
end if
▷ **Candidate fragment discovery**
Construct a spectral embedding from $\mathbf{Z}^{\mathrm{final}}$ and overcluster the regions into $\mathcal{F} = \{\mathcal{F}_j\}_{j=1}^{M}$
Compute fragment prototypes $\{\mathbf{o}_j\}$ and aggregate $\{r_j^F, u_j^F\}$ from their member regions
▷ **GM-OT: globally coupled fragment assignment**
Initialize provisional assignments $\{h(j)\}$, primitive prototypes $\{\mathbf{p}_k^{(0)}\}_{k=1}^{K}$, and slot capacities
Construct source and target marginals $\mathbf{a}$ and $\mathbf{b}$
repeat
    Compute feature distances and competitive margins $\{d_{jk}, m_{jk}\}$
    Construct the reliability-, risk-, margin-, and redundancy-aware transport cost $\mathbf{C}$
    Solve the entropy-regularized transport problem by Sinkhorn scaling:

$$\mathbf{P}^* \leftarrow \mathrm{Sinkhorn}(\mathbf{C}, \mathbf{a}, \mathbf{b})$$

    Re-estimate primitive prototypes from the transport-weighted fragment assignments
until the transport refinement terminates
Assign each fragment to a primitive slot: $\widehat{y}_j^F \leftarrow \arg\max_k P_{jk}^*$
Propagate fragment assignments to neural regions and points, producing $\widehat{\mathbf{y}}$
Train S_θ using $\widehat{\mathbf{y}}$ as hard pseudo-labels
return $\widehat{\mathbf{y}}$ and S_θ

F Additional Experimental Results

F.1 Evaluation on ScanNet

Tables 11 and 12 present the per-category results on the ScanNet validation and hidden test splits. IFOT-SSeg achieves the strongest overall performance among the reported unsupervised methods, with recurring advantages over LogoSP on walls, chairs, desks, shower curtains, sinks, and bathtubs across both splits. The improvements on shower curtains and sinks are particularly informative, as several baselines obtain little or no overlap with these categories. Together with stronger recognition

of desks and bathtubs, these results indicate broader semantic recovery beyond the floor category, for which several existing methods already perform well.

The similar distribution of category-level gains across the validation and hidden test splits supports the generalization of the proposed framework. Its benefits extend to both structural surfaces and individual objects, consistent with the goal of refining mixed region representations and preserving distinct semantic structures during fragment assignment. Some categories, including pictures and refrigerators, remain challenging, indicating that the benefits depend on the available discriminative information. Overall, the results suggest that latent factor induction and globally coordinated assignment help recover semantic distinctions that are insufficiently captured by the compared clustering-based pipelines.

The detailed visualizations in Fig. 5 further illustrate these category-level differences. In the bathroom scenes, the highlighted sink and curtain regions show closer agreement with the ground truth, with fewer incorrect labels extending across neighboring structures. In the meeting-room scene, chairs surrounding the tables are assigned more consistent semantic labels, while the table surfaces remain distinct. These examples illustrate improved semantic separation between adjacent objects and more coherent predictions within individual regions, complementing the quantitative results.

Table 11: Per-category quantitative results on the validation split of ScanNet dataset.

	OA(%)	mAcc(%)	mIoU(%)	wall	floor	cab.	bed.	chair.	sofa.	table.	door.	wind.	books.	pic.	counter.	desk.	curtain.	fridge.	shower.	toilet.	sink.	bathtub.	otherf.
K-means	10.1	10.0	3.4	9.0	9.8	3.2	2.9	5.5	3.3	4.3	3.5	5.5	3.3	**2.6**	0.8	2.9	4.3	0.8	0.8	0.7	0.3	0.9	4.0
IIC	27.7	6.1	2.9	25.3	20.5	0.6	0.3	3.7	0.4	1.3	1.3	1.1	1.9	0.2	0.1	0.6	0.3	0.4	0	0	0	0.2	0.5
PiCIE	20.4	16.5	7.6	14.7	24.5	6.3	5.2	18.0	8.4	33.2	6.7	4.8	9.3	2.1	0.1	2.7	8.0	1.1	2.1	0	0.5	5.0	5.0
GrowSP	57.3	44.2	25.4	40.7	89.8	**24.0**	47.2	45.5	43.0	39.4	14.1	20.0	53.5	0.1	5.4	13.3	8.4	**2.1**	11.3	20.6	19.4	0	9.8
PointDC	62.4	38.8	26.0	59.1	**94.0**	22.0	43.2	30.4	35.9	38.3	14.3	37.4	44.4	1.2	2.4	2.3	**39.4**	2.0	0	38.6	0	2.2	12.7
PointDC-DINOv2	64.7	45.0	29.6	57.0	86.0	18.5	60.6	60.1	46.2	46.4	27.0	39.0	54.2	0.7	25.0	18.1	22.7	0.2	2.8	16.8	0	0	10.4
LogoSP	64.7	50.8	35.8	46.3	86.6	20.7	**66.8**	63.3	**50.9**	**47.1**	**33.8**	41.6	**62.8**	1.0	38.0	10.5	28.6	0.5	0	**46.3**	0	42.3	**29.6**
Ours	**72.5**	**55.5**	**41.3**	**66.0**	89.5	21.9	63.3	**66.0**	50.0	45.7	28.5	**42.2**	54.4	0	**44.4**	**29.8**	30.3	1.0	**45.5**	44.4	**22.3**	**66.9**	13.5

Table 12: Per-category quantitative results on the hidden test split of ScanNet dataset.

		mIoU(%)	wall.	floor.	cab.	bed.	chair.	sofa.	table.	door.	wind.	books.	pic.	counter.	desk.	curtain.	fridge.	shower.	toilet.	sink.	bathtub.	otherf.
Sup.	PointNet++	33.9	52.3	67.7	25.6	47.8	36.0	34.6	23.2	26.1	25.2	45.8	11.7	25.0	27.8	24.7	18.3	14.5	54.8	36.4	58.4	18.3
	DGCNN	44.6	72.3	93.7	36.6	62.3	65.1	57.7	44.5	33.0	39.4	46.3	12.6	31.0	34.9	38.9	28.5	22.4	62.5	35.0	47.4	27.1
	PointCNN	45.8	70.9	94.4	32.1	61.1	71.5	54.5	45.6	31.9	47.5	35.6	16.4	29.9	32.8	37.6	21.6	22.9	75.5	48.4	57.7	28.5
	SparseConv	72.5	86.5	95.5	72.1	82.1	86.9	82.3	62.8	61.4	68.3	84.6	32.5	53.3	60.3	75.4	71.0	87.0	93.4	72.4	64.7	57.2
Unsup.	GrowSP	26.9	32.8	**89.6**	15.2	62.9	55.3	38.9	32.0	14.4	23.0	**59.9**	0.0	12.5	11.4	6.1	1.2	9.3	43.9	14.0	0.0	16.5
	LogoSP	32.7	41.4	87.1	18.1	**68.4**	56.2	49.9	**39.6**	**30.2**	**48.7**	49.2	**0.1**	**29.1**	7.3	33.4	0.0	0.0	**54.3**	0.0	21.1	**19.3**
	Ours	**39.8**	**65.1**	89.4	**18.4**	63.5	**61.7**	**54.4**	35.5	30.0	46.4	33.1	0.0	28.8	**34.6**	**43.6**	**1.4**	**61.2**	46.9	**17.3**	**53.7**	10.5

F.2 EVALUATION ON S3DIS

Tables 13– 18 provide per-category results across the six S3DIS areas. A recurring strength of IFOT-SSeg is its ability to recover categories that are poorly represented by several unsupervised baselines. This is particularly evident for boards in Areas 1, 3, 5, and 6, where the proposed method produces substantially stronger segmentation. Windows in Areas 1, 2, and 6 and doors in Areas 1, 2, 5, and 6 also show repeated advantages. These gains accompany competitive ceiling and floor recognition, suggesting that the framework preserves major scene structures while improving the recovery of additional semantic categories.

The category-level benefits vary across areas, with particularly pronounced gains on sofas in Areas 3 and 5, while columns remain challenging in several areas. This pattern is consistent with the need for region-adaptive representations. Table 19 further summarizes the six-fold results, where IFOT-SSeg achieves the highest mIoU and mAcc among the reported unsupervised methods, with notable advantages on boards, sofas, windows, and doors.

Figure 6 provides complementary visual evidence across different indoor scenes. In the first example, the highlighted wall-mounted board is distinguished from the surrounding wall, consistent with the



Figure 5: Qualitative results on ScanNet. Colors denote semantic classes; highlighted regions show differences between predictions and ground truth.

Table 13: Quantitative results of our method and baselines on the Area-1 of S3DIS dataset.

		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised	PointNet	75.4	74.8	55.0	88.3	93.2	69.2	49.5	37.8	74.5	65.6	41.2	42.5	22.3	35.4	40.9
	PointNet++	76.1	77.9	58.2	90.5	94.4	65.7	38.2	31.9	61.5	66.0	45.3	60.4	41.2	45.8	57.4
	SparseConv	89.0	79.5	72.5	93.6	95.6	76.1	65.9	60.9	60.0	74.2	81.9	85.4	69.2	73.4	33.5
Unsupervised	Kmeans	20.9	24.1	10.1	15.4	17.8	10.5	16.8	1.9	16.0	12.1	9.9	8.1	0.1	6.2	6.7
	IIC	29.2	14.3	8.0	17.0	31.4	25.6	4.3	11.1	0	2.6	1.4	0.7	0	0.2	1.4
	PiCIE	45.7	28.3	19.4	77.2	63.1	24.5	15.8	3.3	4.4	9.6	10.2	14.7	0	9.9	0
	GrowSP	72.9	60.4	45.6	94.2	90.8	52.7	36.7	19.7	33.3	35.8	66.5	**72.6**	13.1	31.2	16.7
	PointDC	58.0	41.5	28.8	88.7	89.5	31.9	1.5	7.1	17.6	12.4	46.4	17.0	0	32.7	0
	PointDC-DINOv2	73.8	55.4	44.0	85.7	**93.7**	58.9	10.1	**20.2**	0	45.0	**70.1**	61.5	**44.0**	39.4	0
	LogoSP	76.9	60.0	48.9	89.0	93.2	**63.2**	27.5	19.2	71.3	39.7	69.7	69.2	0.7	43.6	0
	Ours	**78.7**	**75.6**	**54.8**	**95.4**	88.8	60.5	**49.1**	0.8	**77.5**	**47.2**	60.7	70.7	23.7	**44.9**	**38.6**

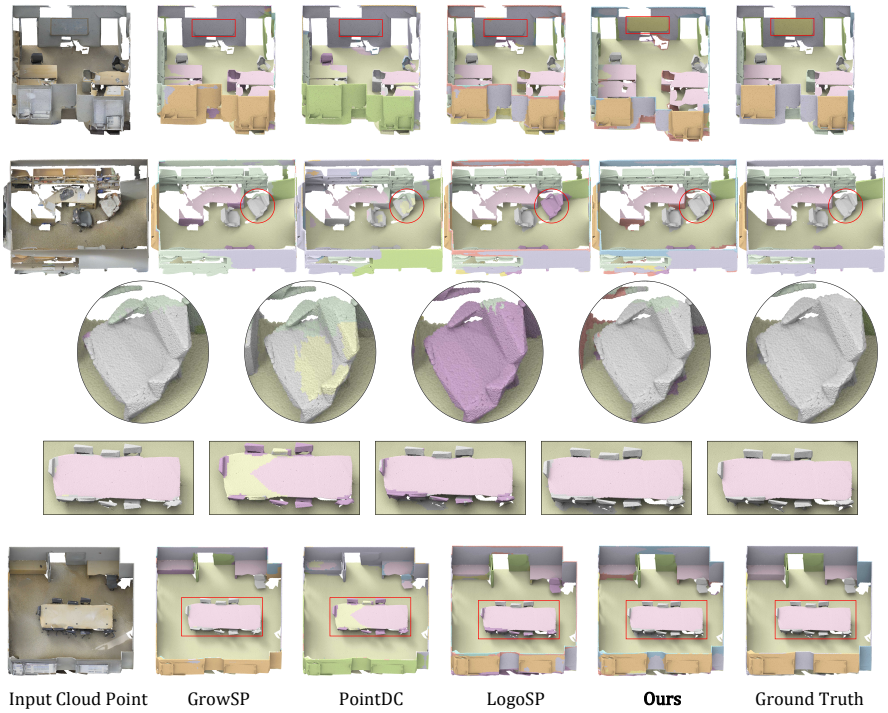


Figure 6: Qualitative results on S3DIS dataset. Colors denote semantic classes; highlighted regions show differences between predictions and ground truth.

recurring gains on the board category. The enlarged sofa region shows more coherent labeling than PointDC and LogoSP, whose predictions contain incorrect labels within the highlighted object. In the meeting-room example, the tabletop is segmented more consistently, with less label mixing across its surface. Together, these examples illustrate improved object-level semantic consistency while preserving distinctions between neighboring scene elements.

Table 14: Quantitative results of our method and baselines on the Area-2 of S3DIS dataset.

		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised	PointNet	72.5	55.5	36.6	79.2	87.4	64.9	14.5	8.2	14.8	39.6	28.8	64.0	7.8	24.4	5.1
	PointNet++	72.1	62.3	39.9	85.8	69.6	71.2	24.9	27.5	32.5	43.6	27.4	51.3	6.0	26.8	12.4
	SparseConv	87.9	69.5	57.3	89.5	93.8	77.0	29.1	32.5	65.5	45.7	67.9	88.8	34.9	54.5	8.2
Unsupervised	Kmeans	17.6	16.6	6.4	16.4	15.6	11.3	3.3	0.9	0.4	6.8	3.7	11.0	1.4	4.6	1.5
	IIC	41.6	16.8	10.6	33.0	43.7	27.6	1.7	0	0	5.6	0.1	13.0	0	2.8	0
	PiCIE	48.3	27.2	17.4	72.4	44.2	39.6	6.2	1.7	0.5	7.7	4.1	20.1	0	7.7	3.6
	GrowSP	79.0	51.8	39.1	85.7	88.2	67.0	**12.0**	**24.8**	0	24.2	51.2	77.1	4.1	24.5	0.2
	PointDC	48.3	34.8	22.5	66.7	50.2	26.1	1.3	0.4	0	15.6	29.8	56.3	0.3	17.6	**5.6**
	PointDC-DINOv2	77.1	50.3	38.1	90.8	**92.0**	57.3	10.2	0.6	34.2	20.3	46.4	**85.6**	0	19.1	0.9
	LogoSP	77.0	52.2	39.4	**92.3**	67.7	**72.0**	7.2	0.7	44.8	32.4	**58.0**	53.3	0.2	44.1	0.5
	Ours	**82.0**	**67.5**	**47.1**	88.5	90.1	71.5	9.9	0.6	**62.04**	**55.9**	50.2	70.2	**16.3**	**47.1**	2.9

Table 15: Quantitative results of our method and baselines on the Area-3 of S3DIS dataset.

		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised	PointNet	78.2	74.9	57.7	90.3	96.9	66.9	55.5	15.1	60.0	67.7	51.8	54.8	27.6	56.0	50.0
	PointNet++	79.8	85.9	65.8	91.4	98.0	68.5	50.1	15.2	74.8	74.7	63.2	70.1	53.6	54.0	76.5
	SparseConv	91.3	86.8	78.6	93.1	96.2	80.4	74.7	63.3	77.2	69.5	80.1	85.5	89.5	80.1	52.5
Unsupervised	Kmeans	21.3	22.1	9.4	20.2	20.6	13.3	5.7	1.3	2.3	14.1	6.8	6.8	3.7	9.7	8.6
	IIC	32.1	15.4	8.4	20.5	25.5	31.4	1.0	6.9	0.2	3.2	1.6	0.3	0	10.6	0
	PiCIE	40.4	29.2	16.2	50.5	49.6	33.7	13.2	3.0	1.8	6.5	8.9	7.5	3.5	16.2	0.4
	GrowSP	74.2	68.4	47.7	92.9	91.7	48.3	**49.3**	15.8	21.1	38.7	60.6	**66.5**	28.5	59.2	0
	PointDC	56.2	40.7	27.2	75.3	91.3	29.7	1.2	2.2	0	11.4	37.9	20.7	9.2	38.4	8.9
	PointDC-DINOv2	70.5	52.4	39.3	86.2	93.4	48.8	0	14.5	38.6	28.2	**68.7**	55.0	0	37.7	0
	LogoSP	**79.8**	62.7	48.9	90.0	**94.3**	**65.9**	16.0	**18.6**	**67.8**	**45.5**	59.2	56.3	3.6	**69.4**	0
	Ours	75.5	**77.5**	**53.9**	**93.5**	92.6	46.6	32.0	0.1	67.4	30.7	63.1	65.4	**55.1**	69.0	**31.0**

Table 16: Quantitative results of our method and baselines on the Area-4 of S3DIS dataset.

		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised	PointNet	73.0	58.6	41.6	81.3	95.7	68.4	1.3	22.4	29.0	44.8	39.3	42.5	17.6	36.6	20.1
	PointNet++	74.8	66.4	47.7	85.5	96.1	69.9	4.4	23.8	27.0	50.5	44.9	54.0	35.6	38.4	43.8
	SparseConv	88.3	76.2	65.5	93.0	94.9	78.2	53.3	57.9	43.4	59.1	69.4	76.6	55.1	73.8	30.9
Unsupervised	Kmeans	17.9	19.9	7.8	18.6	18.2	10.6	0.9	3.8	5.2	11.7	5.8	7.4	2.4	8.7	0.4
	IIC	33.0	13.5	8.2	14.9	25.9	35.1	0	1.1	1.4	3.7	4.1	0.9	0	10.8	0
	PiCIE	43.2	29.4	17.8	62.2	72.7	22.6	2.5	3.4	3.5	8.8	4.1	17.4	0	15.5	0.7
	GrowSP	76.0	**59.8**	42.8	90.6	91.5	64.4	**15.9**	**7.6**	27.4	31.5	52.0	67.4	16.8	48.5	0
	PointDC	54.0	35.3	25.2	87.9	86.8	24.7	0	4.2	12.2	18.7	32.6	17.6	1.9	15.6	0
	PointDC-DINOv2	73.0	46.5	39.8	89.4	91.9	58.2	0	0.7	25.3	19.0	52.4	53.8	0.7	59.0	0
	LogoSP	**80.8**	54.4	43.5	**92.5**	**93.8**	**73.9**	4.1	0.1	34.8	44.6	54.0	64.9	**69.2**	**73.4**	**33.5**
	Ours	76.8	59.4	**48.0**	92.3	93.2	62.7	0.00	6.4	**36.6**	**51.4**	**54.6**	**69.3**	52.9	56.3	0

Table 17: Quantitative results of our method and baselines on the Area-5 of S3DIS dataset.

		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised	PointNet	77.5	59.1	44.6	85.2	97.4	72.3	0.1	10.6	54.9	18.5	48.4	39.5	12.4	55.5	40.2
	PointNet++	77.5	62.6	50.1	83.1	97.2	66.4	0	8.1	55.6	15.2	60.4	64.5	36.6	58.3	55.7
	SparseConv	88.4	69.2	60.8	92.6	95.9	77.2	0.1	36.7	37.6	59.8	77.2	83.9	59.7	78.5	30.4
Unsupervised	Kmeans	21.4	21.2	8.7	18.7	18.0	16.7	**0.2**	2.5	12.0	5.7	8.7	5.6	0	13.6	**2.3**
	IIC	28.5	12.5	6.4	6.1	19.8	27.9	0	2.1	0.1	3.4	7.9	0.4	0	8.6	0
	PiCIE	61.6	25.8	17.9	65.7	61.4	58.4	0	0.3	2.2	1.7	12.1	0	0	12.4	0
	GrowSP	78.4	57.2	44.5	90.5	90.1	66.7	0	**14.8**	27.6	45.6	59.4	71.9	**10.7**	56.0	0.2
	PointDC	55.5	35.1	23.9	84.4	84.3	30.2	0	1.8	12.2	7.1	24.6	6.9	5.4	29.7	0.7
	PointDC-DINOv2	75.7	48.7	40.2	87.7	89.5	59.2	0	0.8	25.8	26.3	**62.0**	68.3	1.5	61.0	0.5
	LogoSP	**82.8**	55.9	46.5	**92.9**	**95.4**	**73.2**	0	3.3	**57.8**	35.9	55.5	**74.6**	1.9	67.3	0.3
	Ours	81.5	**66.6**	**52.8**	92.8	90.8	70.0	0	7.4	56.7	**49.1**	53.9	69.4	**42.7**	**69.7**	**31.5**

Table 18: Quantitative results of our method and baselines on the Area-6 of S3DIS dataset.

		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised	PointNet	79.0	79.6	60.9	85.7	96.5	71.8	59.4	47.4	67.4	74.3	56.2	48.9	20.9	50.0	52.5
	PointNet++	82.0	89.3	69.0	87.5	96.3	76.8	66.4	54.4	72.1	77.4	64.3	66.5	43.7	51.8	70.2
	SparseConv	91.6	87.3	80.5	97.4	95.0	83.4	83.0	75.1	81.1	74.9	81.3	84.3	79.0	80.7	61.4
Unsupervised	Kmeans	21.0	25.0	10.4	18.6	17.6	8.9	11.3	0.6	14.8	17.6	12.0	8.7	0.3	7.8	**6.2**
	IIC	32.5	15.9	9.2	21.9	33.8	29.1	3.1	15.2	0	2.7	0.7	0	0	1.5	1.8
	PiCIE	39.3	28.5	17.8	56.9	61.7	18.6	20.5	4.2	6.0	8.7	14.7	15.9	1.1	5.7	0
	GrowSP	75.6	58.5	47.6	89.4	88.0	57.7	**70.6**	2.0	32.4	36.7	63.2	69.8	1.5	58.9	0.2
	PointDC	62.4	38.7	28.6	85.8	85.6	43.8	5.0	16.5	8.8	10.7	41.7	12.5	0	33.0	0
	PointDC-DINOv2	76.4	55.5	46.4	90.3	91.4	61.7	0	19.7	63.1	33.6	67.9	65.7	1.4	62.5	0
	LogoSP	77.9	62.9	50.6	84.6	**92.7**	64.0	25.7	23.9	65.9	38.8	68.9	**72.2**	2.5	**68.4**	0
	Ours	**78.6**	**70.3**	**53.6**	**92.5**	89.7	58.7	55.5	3.5	**68.9**	**50.5**	65.2	63.8	0	60.1	**35.4**

Table 19: Quantitative results of 6-fold cross validation on S3DIS dataset.

	Method	OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised	PointNet	75.9	67.1	49.4	85.0	94.5	68.9	30.0	23.6	50.1	51.8	44.3	48.7	18.1	43.0	34.8
	PointNet++	77.1	74.1	55.1	85.7	91.6	69.8	36.0	28.0	58.7	57.4	47.4	61.8	39.1	44.1	61.2
	SparseConv	89.4	78.1	69.2	94.6	95.5	78.6	51.8	55.8	60.6	63.0	76.0	84.3	65.8	73.5	39.4
Unsupervised	K-means	20.0	21.5	8.8	17.9	17.9	11.6	6.5	1.8	8.2	11.1	7.9	7.9	1.5	8.4	4.8
	IIC	32.8	14.7	8.5	18.9	30.0	29.3	1.7	7.2	0.4	3.4	2.8	2.5	0.0	5.5	0.6
	PiCIE	46.4	28.1	17.8	63.6	58.6	33.3	9.0	2.6	3.2	7.6	9.7	12.4	0.9	11.5	0.9
	GrowSP	76.0	59.4	44.6	**90.7**	89.9	60.2	**30.6**	**14.9**	24.0	35.6	58.4	**70.6**	12.5	44.9	3.5
	PointDC	55.7	37.7	26.0	81.5	81.5	31.4	1.5	6.4	8.8	12.2	35.8	20.8	2.6	27.5	2.4
	PointDC-DINOv2	74.4	51.5	41.3	88.8	**91.8**	57.9	3.4	10.4	27.5	28.3	**61.3**	64.8	13.8	43.9	0.3
	LogoSP	**79.2**	58.0	46.3	90.2	89.5	**68.7**	13.4	10.9	57.1	39.5	60.9	65.1	13.0	**60.4**	5.7
	Ours	78.9	**69.5**	**51.7**	**92.5**	90.9	61.7	24.4	3.1	**61.5**	**47.5**	58.0	68.1	**31.8**	57.9	**23.2**

F.3 EVALUATION ON NUSCENES

Table 20 presents the per-category results on nuScenes. IFOT-SSeg achieves the highest mIoU and mAcc among the compared methods, with particularly pronounced gains on buses, sidewalks, and terrain. The strong recovery of buses, which are poorly recognized by the baselines, highlights the ability of the proposed framework to capture semantic distinctions that existing unsupervised predictions often miss. These results support its applicability to sparse outdoor point clouds.

Figure 7 provides complementary qualitative comparisons. In the highlighted vehicle regions, our predictions show closer agreement with the ground-truth semantic labels, illustrating improved discrimination between vehicle categories in the displayed scenes.

Table 20: Per-category quantitative results on the validation split of nuScenes dataset.

Method	OA(%)	mAcc(%)	mIoU(%)	barrier	bicycle	bus	car	construction vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	drivable surface	other flat	sidewalk	terrain	manmade	vegetation
GrowSP	39.2	17.5	10.2	7.5	0	0.4	42.9	0.1	0	0.6	0	0.7	1.4	48.4	0.8	6.5	13.1	21.4	19.7
PointDC	**56.8**	29.4	17.7	11.6	0	0.5	63.1	**0.3**	0	4.4	0	1.2	26.4	70.1	0.1	7.1	19.3	21.1	**58.1**
PointDC-DINOv2	51.8	28.6	18.2	**17.0**	0	0.2	58.4	0.2	0	1.5	0	**1.6**	**43.3**	**71.8**	0	8.3	19.5	17.6	51.8
LogoSP	54.8	29.2	20.1	16.6	0	0.7	**70.2**	0.2	**0.2**	**33.6**	0	0.3	38.4	59.4	0.4	8.0	10.7	**33.0**	49.3
Ours	54.7	**37.3**	**22.7**	13.3	0	**46.9**	58.8	**0.3**	0	18.7	0	0.1	34.4	65.9	**0.8**	**13.8**	**35.4**	22.3	52.1

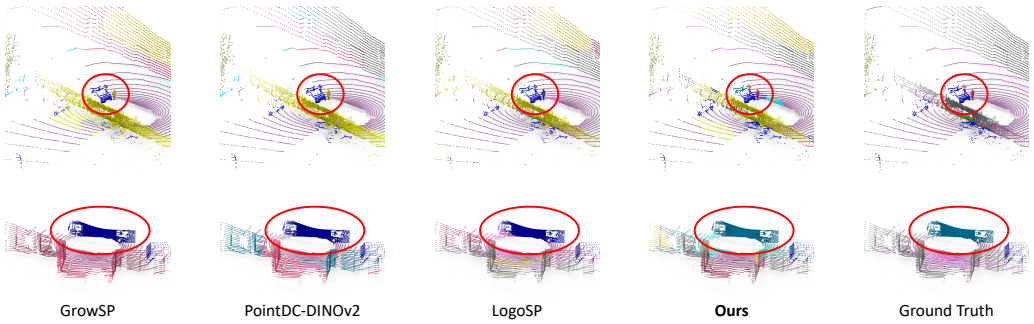


Figure 7: Qualitative results on nuScenes dataset. Colors denote semantic classes; highlighted regions show differences between predictions and ground truth.