RESOLVING LONG-TAIL AMBIGUITY IN UNSUPERVISED 3D POINT CLOUD SEGMENTATION WITH LANGUAGE PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing approaches for unsupervised 3D point cloud segmentation predominantly rely on a *purely visual* similarity-based learning-by-clustering paradigm, which suffers from a fundamental limitation: **long-tail ambiguity**. In such a paradigm, features of minor classes are consistently absorbed by dominant clusters, leading to severely imbalanced predictions. To address this issue, we propose **LangTail**, a language-guided hierarchical learning framework that leverages the balanced world knowledge encoded in language models to mitigate long-tail ambiguity in unsupervised 3D segmentation. The key idea is to establish multi-level associations between language-derived semantic priors and visually underrepresented minor classes, thereby compensating for the biased attention of purely visual clustering toward dominant classes. Specifically, LangTail first constructs an entity-level semantic prior from language models, capturing balanced and fine-grained world knowledge across categories. These priors are injected into a hierarchical clustering framework via contrastive alignment. This guides multi-granularity semantic structure formation and prevents minor classes from being absorbed by dominant clusters, yielding more discriminative representations for underrepresented categories. Extensive experiments on ScanNet-v2, S3DIS, and nuScenes demonstrate that LangTail consistently outperforms existing methods by significant margins, *i.e.*, +13.5, +12.9, and +8.9 mIoU, respectively. These results demonstrate the effectiveness of language priors in improving the representation of minority classes in 3D point clouds.

1 INTRODUCTION

3D point cloud semantic segmentation is a fundamental task for a wide range of applications, including robotic planning Krüsi et al. (2017), autonomous driving Cui et al. (2021), and embodied AI Ze et al. (2024); Huang et al. (2026). While recent fully supervised methods have achieved remarkable progress, they rely heavily on large-scale datasets with dense 3D annotations Hu et al. (2020); Wu et al. (2024); Kolodiazhnyi et al. (2024), which are costly and labor-intensive to obtain. More importantly, in many real-world scenarios such as autonomous robotic systems, manual annotation is often impractical or even infeasible Behley et al. (2019); Hu et al. (2021). To alleviate this dependency, a line of unsupervised methods has been proposed, including GrowSP Zhang et al. (2023), PointDC Chen et al. (2023d), and LogoSP Zhang et al. (2025). These approaches typically follow a learning-by-clustering paradigm, where pseudo-labels are generated based on *purely visual* features and iteratively refined to train the network.

Despite their advances, existing methods suffer from a fundamental limitation: **long-tail ambiguity**. As shown in Fig. 1, while dominant classes are segmented reasonably well, tail classes are consistently misclassified and often achieve near-zero accuracy. This limitation arises because *purely visual similarity-based clustering inherently favors majority patterns*, which appear more frequently in the data, causing features of rare categories to be absorbed into dominant clusters. As a result, the learned representations are biased toward frequent classes, leading to imbalanced and semantically inconsistent predictions.

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.

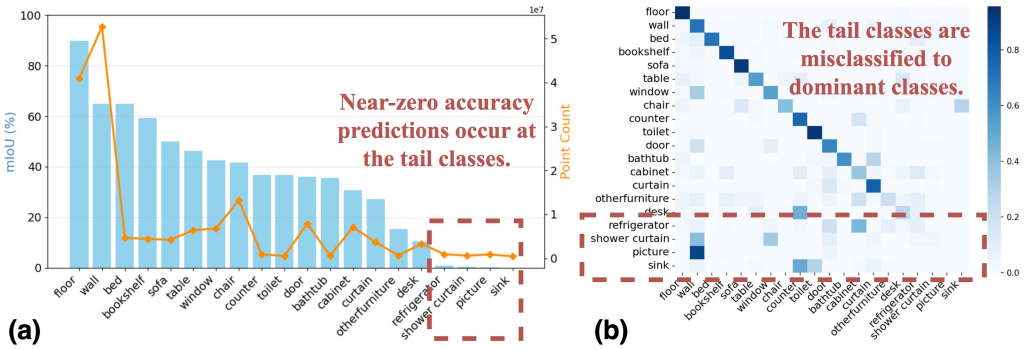


Figure 1: Analysis of the **Long-tail ambiguity** problem in unsupervised 3D semantic segmentation. (a) visualizes the per-category mIoU results (in blue) and the per-category point counts (in orange). The long-tail amiguity problem can be observed, where near-zero predictions occur at the tail class. (b) visualizes the confusion matrix between the predicted results and the ground truth under each category. It can be observed that the tail classes are misclassified to dominant classes.

In contrast, recent advances in vision-language models have shown that language models trained on large-scale corpora encode rich and structured world knowledge Guo et al. (2025); Seed (2026); Yang et al. (2025); Bai et al. (2025). Unlike purely visual representations learned from a limited point cloud dataset, such knowledge captures diverse semantic concepts and their relationships across a wide range of categories Radford et al. (2021). For example, language models can distinguish fine-grained semantic entities (*e.g.*, *shower curtain* vs. *curtain*) and understand their contextual usage, even when their visual appearances are highly similar. Moreover, this knowledge is inherently more balanced across categories, as it is learned from large-scale text corpora rather than the skewed distribution of a specific 3D dataset. As a result, **we claim that language priors provide a complementary and less biased source of 3D semantic supervision**, which is particularly beneficial for modeling tail classes that are underrepresented in visual data.

On top of these observations, we propose **LangTail**, a language-guided hierarchical learning framework for resolving long-tail ambiguity in unsupervised 3D point cloud segmentation. The key idea is to leverage the balanced world knowledge encoded in language models to counteract the bias of purely visual clustering. Concretely, LangTail constructs an entity-level semantic prior using vision-language and segmentation models, and aggregates it into a global semantic bank. This prior is injected into the 3D feature space via contrastive alignment, enabling consistent associations between language-derived entities and 3D features. Built upon this, we present a hierarchical clustering framework over superpoints with a dual-branch design that captures both local and global structures, producing multi-granularity pseudo-labels. Joint optimization with language-guided alignment encourages semantically consistent clustering and prevents tail classes from being absorbed by dominant categories.

We conduct extensive experiments on popular benchmarks, *i.e.*, ScanNet-v2, S3DIS, and nuScenes, and LangTail significantly outperforms existing methods by +13.5, +12.9, and +8.9 mIoU, respectively. These results highlight the effectiveness of language priors in mitigating long-tail ambiguity and improving the segmentation of tail classes in 3D point clouds.

Our main contributions in this work are summarized as follows:

- We are *the first* to leverage language priors to address the long-tail ambiguity problem in unsupervised 3D point cloud segmentation, providing a new perspective for mitigating clustering bias toward dominant classes.
- We propose LangTail, a language-guided hierarchical learning framework that constructs an entity-level semantic prior from vision-language models and injects it into the clustering process via contrastive alignment, enabling balanced and semantically consistent representation learning.
- We develop a dual-branch hierarchical clustering strategy that captures both local geometric consistency and global relational structure, generating multi-granularity pseudo-labels to improve clustering robustness.

2 Related Work

Unsupervised 3D Semantic Segmentation.

Directly adapting 2D unsupervised methods Cho et al. (2021); Ji et al. (2019); Caron et al. (2018) to 3D fails, primarily due to the inherent challenges of point sparsity and occlusion in 2D images. While self-supervised pretraining Xie et al. (2020); Hou et al. (2021) captures discriminative geometry, it fundamentally focuses on low-level representation and lacks inherent high-level semantic understanding. Recent unsupervised 3D segmentation methods Chen et al. (2023d); Liu et al. (2024); Zhang et al. (2023) bypass direct point-wise clustering by adopting superpoints or super-voxels guided iterative self-training. These approaches progressively refine pseudo-labels to distill low-level geometric coherence into high-level semantic discriminability. To mitigate the noise inherently associated with pseudo-label refinement, subsequent works introduce robust clustering mechanisms. LogoSP Zhang et al. (2025) constructs a global superpoint graph and leverages Graph Fourier Transform for frequency-domain clustering, while P-SLCR Zhan et al. (2026) employs a dynamic dual-prototype library to separate high-confidence points from uncertain ones, thereby filtering unreliable supervision signals. GrowSP++ Zhang et al. (2026) uses a dual progressive growing strategy on superpoints and semantic primitives, evolving supervision from local geometry to global semantics to naturally filter early-stage errors. However, these methods ignore minor class ambiguity. Class imbalance and sparsity, coupled with absent semantic priors, leave rare features vulnerable to collapse Van Gansbeke et al. (2020); Hamilton et al. (2022), as distance-based clustering inevitably absorbs them into dominant centroids. To address this, our hierarchical learning framework injects language priors into self-supervised learning, explicitly preserving long-tail representations during optimization.

Cross-modal Priors Assisted 3D Understanding. Cross-modal priors extracted from Vision-Language Models (VLMs) Radford et al. (2021) have emerged as a pivotal resource for advancing 3D understanding. By transferring the rich semantic knowledge encoded in image-text corpora to 3D representations Xu et al. (2024b; 2026; 2021a;b); Chen et al. (2023c); Xu et al. (2023; 2024a), these priors effectively assist 3D models in overcoming inherent data scarcity and weak semantic grounding.

Recent works transfer cross-modal priors via multi-view projection Zhu et al. (2023), feature distillation Chen et al. (2023d); Peng et al. (2023), and triplet alignment Xue et al. (2023); Liu et al. (2023), enabling open-vocabulary and language-grounded 3D perception Thengane et al. (2025). Yet, these paradigms overlook minor-class ambiguity in unsupervised clustering. Distance-based pseudo-labeling inherently marginalizes underrepresented categories, while open-vocabulary methods Peng et al. (2023); Yang et al. (2024) rely on manual prompts or fine-tuning, leaving long-tail bias unresolved without any 3D manual annotations. In contrast, our method constructs an offline entity-level semantic library by synergizing pre-extracted 3D entities with vision-language priors. By injecting these cross-modal priors into a contrastive learning pipeline, the learning of long-tail categories can be guaranteed.

3 LangTail

To handle the long-tail ambiguity problem in unsupervised 3D point cloud segmentation, we propose LangTail, a language-guided hierarchical learning framework without requiring any manual annotations. As illustrated in Fig. 2, LangTail follows an alternating learning-by-clustering pipeline consisting of three complementary branches. The Entity Branch (Sec. 3.1) first constructs an entity-level semantic bank with 2D foundation models (i.e. VLM Guo et al. (2025); Seed (2026); Yang et al. (2025), SAM3 Carion et al. (2026) and clip Radford et al. (2021)) offline, providing balanced language-guided semantic supervision for 3D representation learning. Meanwhile, a cross-modal distillation warm-up provides an initial semantic representation for the 3D backbone. During training, the current 3D features are fed into the Local and Global Branches (Sec. 3.2) , which generate hierarchical pseudo-labels and cluster prototypes from local feature structures and global relationships, respectively. The resulting prototypes initialize the corresponding classifier weights, while the pseudo-labels together with the fixed entity bank supervise the joint optimization of the 3D backbone and classifiers. At the beginning of each subsequent epoch, the updated backbone is used to re-extract features and regenerate the hierarchical pseudo-labels and prototypes, forming an iterative cycle of

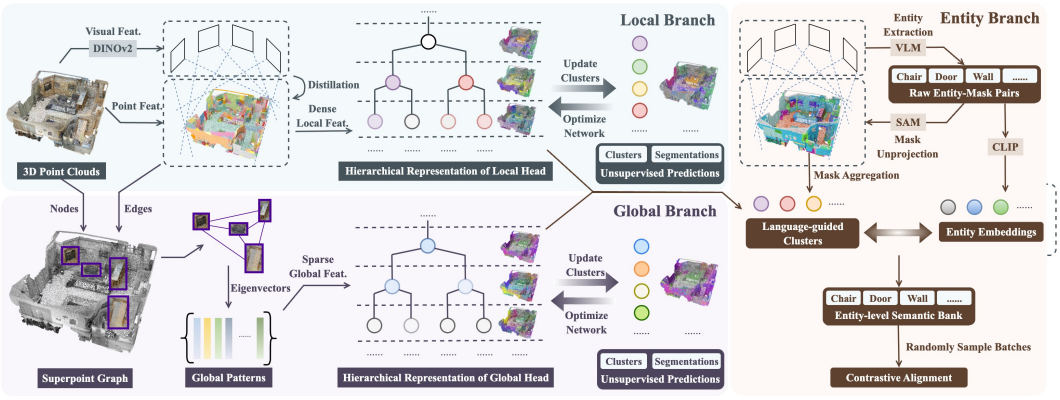


Figure 2: Overview of our LangTail. The Entity Branch inject balanced language priors into the 3D feature space via contrastive alignment. The Local and Global Branches respectively utilize hierarchical learning-by-clustering to excavate the meaningful semantics from local regions and global patterns.

pseudo-label refinement and network optimization. The details of this pipeline are presented in Sec. 3.3.

Problem Statement. Given a dataset with M point clouds $\{\boldsymbol{P}^1 \cdots \boldsymbol{P}^i \cdots \boldsymbol{P}^M\}$, the i-th point cloud $\boldsymbol{P}^i \in \mathbb{R}^{N_i \times 6}$ represents a scene point cloud with N_i points of 6 dimensions including the XYZ coordinates and RGB intensities. For each point cloud $\boldsymbol{P}^i$, there are O RGB images $\{\boldsymbol{I}_1^i \cdots \boldsymbol{I}_o^i \cdots \boldsymbol{I}_O^i\}$ associated with the i-th point cloud. The 2D-3D correspondences between pixels and points can be easily built via the provided depth map and camera parameters in the 3D datasets. Our goal is to discover a set of semantically meaningful classes belonging to c categories for any n-th point $\boldsymbol{P}_n^i$ in the point cloud.

Preliminary. We begin with the preliminaries of prior works Zhang et al. (2023); Chen et al. (2023d); Zhang et al. (2025; 2026) in a learning-by-clustering pipeline. The feature extractor of 3D point cloud is denoted as f_θ parameterized by θ. The extracted feature is noted as $f_\theta(\boldsymbol{P}^i) \in \mathbb{R}^{N_i \times D}$, where D is the feature dimension. The extracted features are iteratively clustered to obtain pseudo-labels, which are then used to optimize f_θ with a segmentation loss. In LangTail, we extend this paradigm with hierarchical local/global clustering and language-guided entity alignment.

3.1 Entity Branch

To inject balanced semantic priors from language into 3D representation, we build an entity-level semantic bank using vision-language models and foundational segmentation models. The entity-level semantic bank contains a huge amount of entity concepts in text and their corresponding 2D masks. The 3D masks are generated by unprojecting the 2D masks using the 2D-3D correspondence. The aggregated 3D priors and entity-level semantics in the bank are then injected into the 3D feature space via contrastive alignment, bridging the associations between language-derived entities and 3D features.

Raw Multi-Modal Priors. For a point cloud $\boldsymbol{P}^i$ and its associated O images, we feed each image to the VLM (Qwen3) and generate a list of entity categories presented in the image. The entities are listed in the text response of VLM, such as "[Chair, Desk, ...]". Then we separately input the entity-level text descriptions to the foundation segmentation model (SAM3) to generate the corresponding 2D masks. Since one pixel might be covered by several different 2D masks, we filter out these labels with lower confidence from the foundation segmentation model, and retain only one single mask per pixel in the end. The filtered 2D masks are then warped back to 3D space according to the 2D-3D correspondence between point clouds and images. After 2D-to-3D projection, if a 3D point corresponds to multiple masks, we keep the one with the highest SAM3 confidence. Points without any mask are assigned ID $= -1$ and excluded from the entity loss. In this way, we can obtain paired samples comprised of entity-level text representations and the corresponding 3D masks.

Entity-level Semantic Bank. We aim to generate an entity-level semantic bank $\boldsymbol{B} \in \mathbb{R}^{T\times C}$ from the extracted raw multi-modal priors in previous section. T is the total number of extracted entities and C is the feature dimension of the 3D backbone $f_{\boldsymbol{\theta}}$. Note that the feature dimension of the entity feature in the semantic bank should be the same as the 3D backbone rather than an arbitrary setting to conduct contrastive alignment in the following steps. A naive thought is to build the semantic bank with pairwise entity-level text and 3D masks with CLIP embedding directly, but the raw entity-mask pairs might suffer from the vague descriptions (i.e., red bed vs blue bed, table vs desk). These classes are annotated by the models as separate pairs but should not be treated as distinct categories during contrastive learning. Consequently, we need to suppress the noisy and redundant feature representations in the entity-mask pairs before constructing the semantic bank. We initialize $\boldsymbol{F}_m$ with Eq. 1 and then refine it by solving Eq. 2.

Specifically, we first extract the CLIP embedding $\boldsymbol{F}_e \in \mathbb{R}^{T\times 512}$ from the text descriptions of each entity (e.g., "a photo of a chair"). T represents the total number of entities. The 3D mask of point cloud $\boldsymbol{P}^i$ corresponding to the t-th entity text is noted as $\boldsymbol{M}_e^{i,t} \in \mathbb{R}^{N_i}$. We can calculate the entity-level point feature $\boldsymbol{F}_m \in \mathbb{R}^{T\times C}$ by:

$$\boldsymbol{F}_m^t = \frac{1}{M}\sum_i \left(\frac{1}{|\boldsymbol{M}_e^{i,t}|}\sum f_{\boldsymbol{\theta}}(\boldsymbol{P}^i)\cdot \boldsymbol{M}_e^{i,t}\right) \tag{1}$$

where the extracted original feature $f_{\boldsymbol{\theta}}(\boldsymbol{P}^i)$ from point cloud is averaged under the guidance of 3D entity mask $\boldsymbol{M}_e^{i,t}$ to group the entity-level point features. $\boldsymbol{F}_m^t$ represents the t-th of the T entities in the entity-level point feature $\boldsymbol{F}_m$.

To bridge the dimension gap between $\boldsymbol{F}_m \in \mathbb{R}^{T\times C}$ and $\boldsymbol{F}_e \in \mathbb{R}^{T\times 512}$, we further build a simple optimization problem on the entity-level point feature $\boldsymbol{F}_m$:

$$\min_{\boldsymbol{F}_m} \|\boldsymbol{G}(\boldsymbol{F}_m) - \boldsymbol{G}(\boldsymbol{F}_e)\|_2^2 \tag{2}$$

where $\boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{X}\cdot\boldsymbol{X}^T$ is the Gram matrix. We optimize the difference between gram matrices of entity-level point feature $\boldsymbol{F}_m$ and entity-level text feature $\boldsymbol{F}_e$ to inject the language prior to the point features. This objective function can inherit the structure information of 3D features and merge the entity-level semantics guided by language into the 3D representation in a class-agnostic manner. After solving the optimization problem in Eq. 2, we can obtain the entity-level semantic bank $\boldsymbol{B} = \boldsymbol{F}_m^*$. After construction, the entity-level semantic bank is kept fixed throughout the subsequent optimization and serves as a stable semantic reference for entity-level contrastive alignment.

Entity-level Contrastive Alignment. In consideration of the class agnosticism and the huge number of entities in semantic bank, we randomly sample a subset of entities from the semantic bank each time and conduct contrastive alignment between the sampled entities and the extracted feature from 3D backbone network $f_{\boldsymbol{\theta}}$. Specifically, we use an InfoNCE-style contrastive loss to align the point features with the sampled entity prototypes in the bank during training:

$$\mathcal{L}_{\text{entity}} = -w_{c_i}\frac{1}{M}\sum_i \log \frac{\exp\left(f_{\boldsymbol{\theta}}(\boldsymbol{P}^i)\cdot \boldsymbol{b}^+/\tau\right)}{\exp\left(f_{\boldsymbol{\theta}}(\boldsymbol{P}^i)\cdot \boldsymbol{b}^+/\tau\right) + \sum_{j=1}^{N_{\text{neg}}}\exp\left(f_{\boldsymbol{\theta}}(\boldsymbol{P}^i)\cdot \boldsymbol{b}_j^-/\tau\right)} \tag{3}$$

where $\boldsymbol{b}^+$ is the prototype of the sampled batch, $\boldsymbol{b}_j^-$ is the hard negative samples (prototypes of other classes in the same batch). τ is the temperature parameter. w_{c_i} is the balance weight for category c_i, calculated based on the inverse square root of the category frequency within the training batch: $w_{c_i} = 1/\sqrt{n_{c_i}}$. n_{c_i} is the number of points assigned to entity c_i in the current batch, counted from the projected masks without using any ground-truth labels. This loss guides the alignment of 3D features with the semantic space from pretraining, thereby achieving stronger semantic discrimination capabilities.

3.2 Local Branch and Global Branch

Cross-Modal Distillation Warmup. Following previous works Chen et al. (2023d); Zhang et al. (2025), we first extract pixel-level features with a pretrained visual foundation model (DINO-v2). The pixel-level features are then mapped to the corresponding 3D points via the 2D-3D correspondence between images and point clouds. When multiple pixels map to the same 3D point, we average their DINO features. As a warming up strategy, we distill the semantic information from these unprojected 2D features to 3D features with a cosine similarity loss.

Local Branch. To model the inherent hierarchical structure between different semantic concepts, we perform Ward-based hierarchical clustering directly on the grouped superpoints of 3D features. Specifically, we can construct a hierarchical tree from the bottom to the top and truncate it at predefined level and number of clusters. In default, we set the set of cluster numbers to $\mathcal{K} = \{120, 80, K_{\text{prim}}\}$, where K_{prim} is the number of target semantic primitives.

For each granularity $k \in \mathcal{K}$, we compute a cluster prototype $\boldsymbol{\mu}_k^{\text{local}} \in \mathbb{R}^{k \times C}$ from the resulting clusters. k is the selected number of clusters of hierarchical tree and C is the feature dimension. These prototypes are used to assign superpoints and obtain the corresponding pseudo-labels, and are also used to initialize the corresponding linear classifier weights, denoted as $\boldsymbol{W}_k^{\text{local}}$. The classifier weights are subsequently optimized together with the 3D backbone during the current training epoch.

The backbone network $f_{\boldsymbol{\theta}}$ is optimized via calculating the cross-entropy loss:

$$\mathcal{L}_{\text{local}} = \sum_{k \in \mathcal{K}} \mathcal{L}_{CE}(g_{\boldsymbol{W}_k^{\text{local}}}(f_\theta(\boldsymbol{P}_j^i)), \hat{y}_j^i) \tag{4}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss and $g_{\boldsymbol{W}_k^{\text{local}}}$ is the linear segmentation head parameterized by the classifier weights $\boldsymbol{W}_k^{\text{local}}$. $\hat{y}_j^i$ is the assigned label towards the corresponding cluster prototypes of the j-th point in the i-th point cloud of the dataset.

Global Branch. Although local branch can effectively group the adjacent regions with similar structure and semantics, they are prone to the limitations of local perspective and struggle to accurately group objects with the same semantics that are spatially separated. To address this issue, Global Branch incorporates spectral graph theory to capture the topological relationships and global contexts of 3D points. Following Zhang et al. (2025), the global relationship among superpoints is constructed as a graph. After constructing and grouping global patterns, we obtain the refined graph fourier basis for each superpoint. More details are in the Appendix C.

Similar to Local Branch, we apply hierarchical clustering to generate the pseudo-labels first and then supervise the 3D backbone f_θ with cross-entropy loss:

$$\mathcal{L}_{\text{global}} = \sum_{k \in \mathcal{K}} \mathcal{L}_{CE}(g_{\boldsymbol{W}_k^{\text{global}}}(f_\theta(\boldsymbol{P}_j^i)), \tilde{y}_j^i) \tag{5}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss and $g_{\boldsymbol{W}_k^{\text{global}}}$ is the linear segmentation head parameterized by the classifier weights $\boldsymbol{W}_k^{\text{global}}$. $\tilde{y}_j^i$ is the assigned label towards the corresponding cluster prototypes of the j-th point in the i-th point cloud of the dataset.

3.3 Iterative optimization of LangTail

Our training procedure alternates between pseudo-label refinement and network optimization. At the begining of each epoch, we first extract the current 3D features with f_θ and generate hierarchical pseudo-labels and cluster prototypes from the Local and Global Branches. The resulting prototypes are used to initialize the corresponding classifier weights $\boldsymbol{W}_k^{\text{local}}$ and $\boldsymbol{W}_k^{\text{global}}$. Together with the fixed entity-level semantic bank, these pseudo-labels provide three supervision signals:

$$\min_{\theta, \{\boldsymbol{W}_k^{\text{local}}, \boldsymbol{W}_k^{\text{global}}\}_{k \in \mathcal{K}}} \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{global}} + \lambda \mathcal{L}_{\text{entity}}, \tag{6}$$

During optimization, the three losses are jointly backpropagated to update f_θ, while $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ additionally update their corresponding classifier weights. At the beginning of the next epoch, the updated features are then reclustered by the Local and Global Branches to produce refined pseudo-labels for the next iteration. This alternating procedure progressively refines the pseudo supervision and the learned 3D representations.

4 Experiments

To evaluate and compare with other methods, we choose two large-scale indoor datasets: (1) **ScanNet-v2** Dai et al. (2017) contains 1,613 scenes, among which 1,201 scenes are used as the training set, 312 scenes as the validation set, and the remaining 100 scenes as the online testing set. (2) **S3DIS** Armeni et al. (2017) consists of six areas, divided into a total of 271 rooms. We select Area 5 as the validation

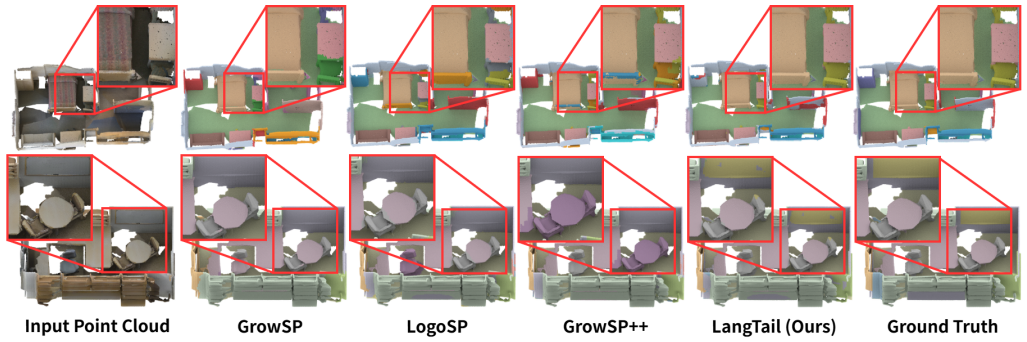


Figure 3: Qualitative comparison on ScanNet (the first row) and S3DIS (the second row) datasets.The red box highlights the differences

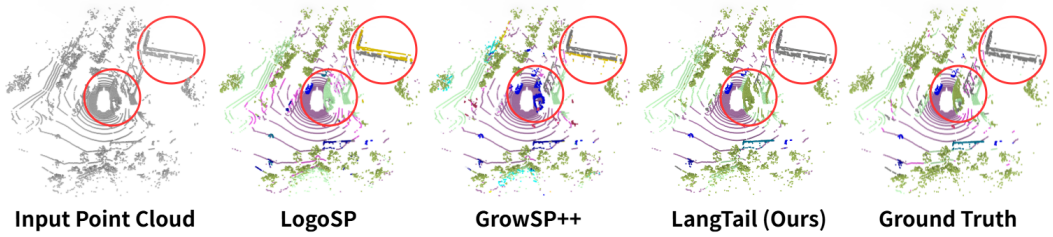


Figure 4: Qualitative comparison of point cloud segmentation on the nuScenes datasets.

set, with the other five areas serving as the training set; and a challenging autonomous driving dataset: (3) **nuScenes** Caesar et al. (2020), as the outdoor dataset, contains 28,130 LiDAR scenes for training and 6018 scenes for validation. Following LogoSP, we choose mean Intersection-over-Union (mIoU), Overall Accuracy (OA), and mean Accuracy (mAcc) across all classes as our metrics. More experiments and detail results are shown in Appendix D.

4.1 Evaluation on S3DIS

S3DIS Armeni et al. (2017) comprises 6 areas and 271 rooms, with each point belonging to one of 12 categories. Following previous methods, we exclude clutter in loss computation or evaluation. We employed 6-fold cross-validation, training on 5 areas and evaluating on the remaining one.

Table 1: Results on ScanNet (Val) and S3DIS datasets.

Method	ScanNet (Val)			S3DIS		
	OA	mAcc	mIoU	OA	mAcc	mIoU
Supervised Methods						
PointNet Qi et al. (2017a)	-	-	-	77.5	59.1	44.6
PointNet++ Qi et al. (2017b)	-	-	-	77.5	62.6	50.1
SparseConv Graham et al. (2018)	-	-	-	88.4	69.2	60.8
Unsupervised Methods						
IIC Ji et al. (2019)	27.7	6.1	2.9	28.5	12.5	6.4
IIC-S Ji et al. (2019)	18.3	6.7	3.4	29.2	13.0	6.8
IIC-PFH Ji et al. (2019)	25.4	6.3	3.4	28.6	16.8	7.9
IIC-S-PFH Ji et al. (2019)	18.9	6.3	3.0	31.2	16.3	9.1
PICIE Cho et al. (2021)	20.4	16.5	7.6	61.6	25.8	17.9
PICIE-S Cho et al. (2021)	35.6	13.7	8.1	49.6	28.9	20.0
PICIE-PFH Cho et al. (2021)	33.1	14.0	8.1	54.0	36.8	24.4
PICIE-S-PFH Cho et al. (2021)	23.6	15.1	7.4	48.4	40.4	25.2
PointDC Chen et al. (2023d)	62.4	38.8	26.0	55.5	35.1	23.9
PointDC-DINOv2 Chen et al. (2023d)	64.7	45.0	29.6	75.7	48.7	40.2
GrowSP Zhang et al. (2023)	57.3	44.2	25.4	78.4	57.2	44.5
LogoSP Zhang et al. (2025)	64.7	50.8	35.8	82.8	55.9	46.5
GrowSP++ Zhang et al. (2026)	70.5	48.8	33.2	78.7	60.1	46.6
LangTail (Ours)	**78.7**	**59.2**	**46.7**	**86.7**	**69.6**	**59.5**

Results & Analysis: As shown in table 1 , our method outperforms all baselines. Compared to the latest method GrowSP++ Zhang et al. (2026), our method achieves improvements of +8.0 in OA, +9.5 in mAcc, and +12.9 in mIoU. Furthermore, we outperform GrowSP++ on all categories, particularly in the board category, where accuracy improved from 0% to 52.4%. As shown in Figure 3, our method not only yields cleaner boundaries but also accurately segments minor classes that baselines completely miss.

4.2 Evaluation on ScanNet

We conduct a comprehensive evaluation of our proposed method on ScanNet Dai et al. (2017), a large-scale real-world dataset of indoor scenes. ScanNet consists of 1,201 scenes in the training set, 312 scenes in the validation set, and 100 scenes in the online test set. Each point in the dataset belongs

to one of 20 categories or an undefined background. Following the approach of GrowSP++ Zhang et al. (2026) and LogoSP Zhang et al. (2025), we also employ the Hungarian matching algorithm to reorder the predicted labels for scoring, and report results separately for the offline validation set and the online hidden test set. The figure shows the qualitative results.

Results & Analysis: As shown in table 1 and table 2 ,we achieved state-of-the-art performance among current unsupervised methods. Compared to the latest method GrowSP++ Zhang et al. (2026), our method improves by +8.2 OA, +10.4 mAcc, and +13.5 mIoU. Our method not only achieves improvements on major categories but also performs exceptionally well on minor categories. For example, in the bathtub category, GrowSP++ achieves 0% accuracy, whereas our method achieves 58.4% accuracy. It demonstrates similarly outstanding performance on other minor categories such as sink, refrigerator, shower curtain etc., illustrating the effectiveness of our method in addressing the long-tail problem. Qualitative comparisons are visualized in Figure 3.

Table 2: Results on ScanNet (Test) dataset.

Method	mIoU
Supervised Methods	
PointNet++ Qi et al. (2017b)	33.9
DGCNN Wang et al. (2019)	44.6
PointCNN Li et al. (2018)	45.8
SparseConv Graham et al. (2018)	72.5
Unsupervised Methods	
GrowSP Zhang et al. (2023)	26.9
LogoSP Zhang et al. (2025)	32.7
GrowSP++ Zhang et al. (2026)	32.3
LangTail (Ours)	**45.3**

4.3 Evaluation on nuScenes

To demonstrate the generalizability of our method, we also conducted evaluation on nuScenes Caesar et al. (2020), an outdoor dataset for autonomous driving. nuScenes is more challenging than the previous two datasets, comprising 28,130 LiDAR scenes for training and 6,018 scenes for validation. Each point belongs to one of 16 categories.

Table 3: Results on nuScenes dataset.

Method	OA	mAcc	mIoU
Unsupervised Methods			
PointDC Chen et al. (2023d)	56.8	29.4	17.7
PointDC-DINOv2 Chen et al. (2023d)	51.8	28.6	18.2
GrowSP Zhang et al. (2023)	39.2	17.5	10.2
LogoSP Zhang et al. (2025)	54.8	29.2	20.1
GrowSP++ Zhang et al. (2026)	55.7	30.4	19.7
LangTail (Ours)	**73.9**	**35.1**	**29.0**

Results & Analysis: As shown in table 3, our method outperforms all baselines and achieves state-of-the-art results. Compared to GrowSP++ Zhang et al. (2026), our method achieves +18.2 in OA, a +4.7 in mAcc, and +9.3 in mIoU, demonstrating the generalizability of our method in outdoor scenarios.The figure 4 shows the qualitative result.

4.4 Generalization to Unseen Datasets

To verify whether the learned semantic features and multi-granularity prototypes possess generalization capabilities, we conducted a cross-dataset evaluation on two indoor datasets, ScanNet Dai et al. (2017) and S3DIS Armeni et al. (2017): (1) **From ScanNet to S3DIS:** We directly use the model trained on ScanNet to extract features from ScanNet. Following a dual-branch hierarchical clustering process, we obtain a total of 440 cluster centroids, each with a granularity of $[120, 80, 20]$. These cluster centroids serve as classifiers for the S3DIS test regions. (2) **From S3DIS to ScanNet:** Similarly, we directly use the model trained on S3DIS to extract features, yielding a total of 424 cluster centroids with branch-specific granularities of $[120, 80, 12]$. These cluster centroids serve as classifiers for the ScanNet test region.

Results & Analysis: As shown in Table 4, our model significantly outperforms all baseline methods in cross-dataset experiments. This validates that our method can also perform effectively on unfamiliar datasets.

4.5 Ablation Study

We conducted ablation experiments on the ScanNet validation set to evaluate the contribution of each component, with the following variants corresponding to the configurations in our table 5 :

(1) To simultaneously capture semantic information across scales, we first integrate local and global branches. **(2)** Recognizing that purely visual supervision is limited in capturing open-world knowledge, we leverage Vision-Language Models (VLMs) and SAM3 to extract 2D masks.These masks are projected onto the 3D point cloud and processed by network g to extract corresponding features.**(3)**

Table 4: Cross-dataset generalization evaluation. **Left (ScanNet Dai et al. (2017) → S3DIS Armeni et al. (2017)):** Models are trained on the ScanNet dataset and tested on varying areas of S3DIS. **Right (S3DIS → ScanNet):** Models are trained on S3DIS and tested on ScanNet. For compactness, "w/o A1" denotes that Area 1 was excluded during training (i.e., trained on Areas 2-6). The best results are highlighted in **bold**.

Method	ScanNet → S3DIS (Tested on Area i)							S3DIS → ScanNet (Trained missing Area i)						
	A-1	A-2	A-3	A-4	A-5	A-6	**Mean**	w/o A1	w/o A2	w/o A3	w/o A4	w/o A5	w/o A6	**Mean**
IIC Ji et al. (2019)	3.7	3.8	3.8	4.0	3.8	3.7	3.8	3.5	3.4	3.7	3.5	3.5	3.6	3.5
PiCIE Cho et al. (2021)	13.5	12.7	13.4	12.8	11.3	13.1	12.8	5.6	5.1	5.0	5.9	6.0	5.5	5.5
GrowSP Zhang et al. (2023)	24.2	21.9	26.1	25.0	23.7	27.9	24.8	16.9	17.8	16.4	16.1	17.1	15.3	16.6
PointDC Chen et al. (2023d)	23.6	20.9	24.6	19.5	20.1	29.7	23.1	10.3	10.1	10.4	8.4	10.0	9.9	9.9
PointDC-DINOv2 Chen et al. (2023d)	33.8	29.6	33.8	31.7	32.2	36.9	33.0	16.7	15.4	16.4	17.6	14.3	14.6	15.8
LogoSP Zhang et al. (2025)	43.8	37.5	47.0	40.7	44.9	47.9	43.6	16.9	16.9	16.8	16.8	16.7	17.0	16.9
GrowSP++ Zhang et al. (2026)	39.5	32.9	41.7	36.0	36.1	43.8	38.3	17.5	16.4	17.2	16.1	17.5	16.6	16.9
LangTail (Ours)	**48.3**	**42.6**	**50.1**	**44.3**	**47.7**	**49.7**	**47.1**	**19.4**	**19.0**	**19.6**	**18.6**	**18.6**	**19.7**	**19.2**

Table 5: Comprehensive ablation study on the ScanNet validation set. We evaluate the impact of core modules (left) and varying hyperparameters (right). The complete setup is shown in the last row.

Variant	Core Modules			Hyperparameters		mIoU (%)
	Dual-Branch	**Hier. Clust.**	**Lang. Mod.**	**Weight** (λ)	**Granularity** (K)	
Component Ablation (Progressive Addition)						
1	✓			-	-	36.0
2	✓	✓		-	[120, 80, 20]	43.8
3	✓	✓	✓	-	[120, 80, 20]	46.7
Hyperparameter Ablation: Entity Loss Weight (λ)						
4	✓	✓	✓	0.7	[120, 80, 20]	46.2
5	✓	✓	✓	0.8	[120, 80, 20]	46.3
6	✓	✓	✓	1.0	[120, 80, 20]	46.5
7	✓	✓	✓	1.1	[120, 80, 20]	45.8
Hyperparameter Ablation: Clustering Granularity (K)						
8	✓	✓	✓	0.9	[120, 40, 20]	45.4
9	✓	✓	✓	0.9	[80, 40, 20]	44.7
Full	✓	✓	✓	**0.9**	**[120, 80, 20]**	**46.7**

Since the unsupervised nature of the task does not allow us to filter the generated vocabulary, the same object may have different entity names (e.g., "blue bed" and "bed").To address this,we incorporate clip-based alignment to group objects of the same class in the feature space,providing reliable supervision signals.**(4)** To better mitigate the long-tail issue, we perform hierarchical clustering in both the local and global branches to extract prototypes at $\{120, 80, 20\}$ granularity levels, which are used to optimize the network.**(5)~(10)** λ represents the weight of the entity loss, and K denotes the clustering granularity.We performed ablation experiments on $\lambda \in \{0.7, 0.8, 0.9, 1.0, 1.1\}$ and on $K \in \{[120, 40, 20], [120, 80, 20], [80, 40, 20]\}$. In the full method,$\lambda$ and K is consistently set to 0.9 and $[120, 80, 20]$,respectively.

Results & Analysis: As shown in Table 5, **(1)** introducing a two-branch structure yields a score of 36.0 in mIoU, representing a slight improvement over LogoSP's 35.8. **(2)** Introducing hierarchical clustering shows a more significant improvement to 43.8 in mIoU, with marked progress in the segmentation of categories. This indicates that hierarchical clustering provides a stronger structural basis for LangTail. **(3)** Building on this improved clustering structure, the introduce of the language module further raises mIoU to 46.7, providing complementary semantic guidance and further alleviating category ambiguity in the long-tail setting. **(4)** *Robustness*: Our method is relatively robust with regard to hyperparameter selection with mIoU remaining stable across different settings.

5 Conclusion and Discussion

LangTail incorporates real-world prior knowledge derived from Visual Language models, enabling it to recognize a broader range of categories. Following a dual-branch hierarchical clustering process, it captures multi-granularity pseudo-labels, thereby offering a novel solution to the long-tail ambiguity problem in unsupervised 3D point cloud segmentation.We conducted extensive experiments across multiple benchmarks, demonstrating the state-of-the-art performance of our method.Nevertheless, our method still has limitations in outdoor settings, which warrant further investigation.

AI USE STATEMENT

In this work, we used generative AI tools to aid in writing, language editing, and polishing the manuscript. We did not use generative AI tools to generate experimental results, fabricate or modify data, perform experiments, or produce the final scientific claims or conclusions of the work. Other required disclosure tasks are not applicable to this work. All AI-assisted text was reviewed and revised by the authors for accuracy, consistency, and appropriateness. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Our implementation and experimental details are provided in Sec.4 and Appendix B. To facilitate reproducibility, we will release the complete training and preprocessing code, configuration files, entity lists, and pre-computed masks upon publication.

REFERENCES

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.

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.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. *Advances in Neural Information Processing Systems*, 36:75896–75910, 2023a.

Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7020–7030. IEEE, 2023b.

Weitao Chen, Hongbin Xu, Zhipeng Zhou, Yang Liu, Baigui Sun, Wenxiong Kang, and Xuansong Xie. Costformer: Cost transformer for cost aggregation in multi-view stereo. *arXiv preprint arXiv:2305.10320*, 2023c.

Zisheng Chen, Hongbin Xu, Weitao Chen, Zhipeng Zhou, Haihong Xiao, Baigui Sun, Xuansong Xie, et al. 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, 2023d.

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.

Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):722–739, 2021.

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.

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.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15587–15597, 2021.

Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11108–11117, 2020.

Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4977–4987, 2021.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026.

Xu Ji, Joao 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.

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

Philipp Krüsi, Paul Furgale, Michael Bosse, and Roland Siegwart. Driving on point clouds: Motion planning, trajectory optimization, and terrain assessment in generic nonplanar environments. *Journal of Field Robotics*, 34(5):940–984, 2017.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

Jiaxu Liu, Zhengdi Yu, Toby P Breckon, and Hubert PH Shum. U3ds3: Unsupervised 3d semantic scene segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3759–3768, 2024.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023.

Aljoša Ošep, Tim Meinhardt, Francesco Ferroni, Neehar Peri, Deva Ramanan, and Laura Leal-Taixé. Better call sal: Towards learning to segment anything in lidar. In *European Conference on Computer Vision*, pp. 71–90. Springer, 2024.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

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 Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

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

Bytedance Seed. Seed1. 8 model card: Towards generalized real-world agency. *arXiv preprint arXiv:2603.20633*, 2026.

Vishal Thengane, Xiatian Zhu, Salim Bouzerdoum, Son Lam Phung, and Yunpeng Li. Foundational models for 3d point clouds: A survey and outlook. *arXiv preprint arXiv:2501.18594*, 2025.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pp. 268–285. Springer, 2020.

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 (tog)*, 38(5): 1–12, 2019.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4840–4851, 2024.

Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pp. 574–591. Springer, 2020.

Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3030–3038, 2021a.

Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6078–6087, 2021b.

Hongbin Xu, Weitao Chen, Yang Liu, Zhipeng Zhou, Haihong Xiao, Baigui Sun, Xuansong Xie, and Wenxiong Kang. Semi-supervised deep multi-view stereo. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4616–4625, 2023.

Hongbin Xu, Weitao Chen, Baigui Sun, Xuansong Xie, and Wenxiong Kang. Robustmvs: Single domain generalized deep multi-view stereo. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9181–9194, 2024a.

Hongbin Xu, Weitao Chen, Zhipeng Zhou, Feng Xiao, Baigui Sun, Mike Zheng Shou, and Wenxiong Kang. Controlrm: Fast and controllable 3d generation via large reconstruction model. *arXiv preprint arXiv:2410.09592*, 2024b.

Hongbin Xu, Chaohui Yu, Feng Xiao, Jiazheng Xing, Hai Ci, Weitao Chen, Fan Wang, and Ming Li. Cyc3d: Fine-grained controllable 3d generation via cycle consistency regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17895–17903, 2026.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1179–1189, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19823–19832, 2024.

Fuyang Yu, Runze Tian, Zhen Wang, Xiaochuan Wang, and Xiaohui Liang. Cus3d: Clip-based unsupervised 3d segmentation via object-level denoise. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

Lixin Zhan, Jiang Jie, Tianjian Zhou, Yukun Du, Yan Zheng, and Xuehu Duan. P-slcr: Unsupervised point cloud semantic segmentation via prototypes structure learning and consistent reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12349–12357, 2026.

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, 2025.

Zihui Zhang, Weisheng Dai, Bing Wang, Bo Li, and Bo Yang. Growsp++: Growing superpoints and primitives for unsupervised 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2639–2650, 2023.

APPENDIX

A BROADER IMAPACTS

Our work significantly advances unsupervised 3D semantic segmentation by enabling accurate, label-free discovery of complex scene structures directly from raw point clouds. It advances 3D scene understanding, offering clear positive societal impacts by eliminating the need for costly manual annotations and lowering the barrier for deploying 3D perception in robotics, autonomous navigation, augmented reality, and urban digital twins. Regarding negative impacts, we note that: (i) LangTail is presented as a research algorithm without direct deployment pathways or decision-making capabilities; (ii) the technology is modality-agnostic and does not process personal/sensitive data; (iii) its unsupervised nature actually reduces privacy risks by avoiding collection of user-specific labels. Any potential misuse would stem from application-level integration such as surveillance, not from the core algorithm itself.

B IMPLEMENTATION DETAILS

Implementation Details. Consistent with previous methods, we employ a sparse convolutional neural network—Res16FPN18, implemented using MinkowskiEngine as the 3D backbone network. The input point cloud is converted into a voxel grid with a voxel size of 5 cm, and the final feature dimension of the network is set to 384. The network is initialized with distilled weights to accelerate convergence. For the training of the segmentation network, we use the AdamW optimizer with a batch size of 8 and train for a total of 200 epochs. The initial learning rate was set to 1×10^{-4}, and a Poly learning rate scheduling strategy (with an exponent of 0.9) was adopted to gradually reduce the learning rate to a minimum of 1×10^{-8}. To ensure computational efficiency, when the number of superpoints in the full dataset is excessive, we randomly sample up to 30,000 superpoints.

Reproducibility Analysis. All operations about VLM and SAM3 are performed offline during the preprocessing stage, and neither training nor inference requires access to external APIs. The offline preprocessing is performed on NVIDIA V100 GPUs, requiring approximately 37.6, 17.87, and 50.97 GPU hours for ScanNet, S3DIS, and nuScenes, respectively. The resulting semantic banks contain 56,799, 13,079, and 487,627 entries for ScanNet, S3DIS, and nuScenes, respectively, with a feature dimension of 384. The subsequent model training is performed on a single NVIDIA GeForce RTX 3090 GPU, requiring approximately 2.6 hours, 1.5 hours, and 15 hours for ScanNet, S3DIS, and nuScenes, respectively.

Evaluation Protocol. During the validation and testing phases, we concatenate the dynamically updated classifier weights from both branches across all granularities, resulting in dataset-specific number of pseudo-class prototypes (440 for ScanNet, 344 for S3DIS, and 352 for nuScenes) pseudo-class prototypes. The network assigns each point to one of these prototypes by calculating feature similarity. Subsequently, we use Hungarian matching to precisely map these pseudo-classes to the true semantic classes, and use this to compute the final evaluation metrics, including mean Intersection-over-Union (mIoU), mean Accuracy (mAcc), and Overall Accuracy (OA).

C GLOBAL SUPERPOINTS GRAPH CONSTRUCTION AND FREQUENCY-DOMAIN PATTERN GROUPING

Global Superpoint Graph Construction Let $\boldsymbol{F} \in \mathbb{R}^{S \times C}$ (where S is the number of superpoints $C = 384$) denote the initial feature matrix of all superpoints.We construct a global graph $\mathcal{G}$ where each node represents a superpoint. The edge weight a_{ij} between the i-th and j-th nodes is defined as follows:

$$a_{ij} = \exp(-\|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2), \tag{7}$$

where $\boldsymbol{f}_i, \boldsymbol{f}_j$ are the corresponding feature vectors from $\boldsymbol{F}$. We then compute the normalized Laplacian matrix $\boldsymbol{L} = \boldsymbol{D}^{-1/2}(\boldsymbol{D} - \boldsymbol{A})\boldsymbol{D}^{-1/2}$, where $\boldsymbol{A}$ is the adjacency matrix and $\boldsymbol{D}$ is the degree matrix. By performing eigen-decomposition on $\boldsymbol{L}$:

$$\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top, \quad \boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s, \ldots, \boldsymbol{u}_S], \tag{8}$$

Table 6: Controlled experiments on the effect of external foundation-model priors.

Method	External Language-guided Priors	OA(%)	mAcc(%)	mIoU(%)
LogoSP	w/o	64.7	50.8	35.8
LogoSP*	w	69.6(+4.9)	52.8	37.2
GrowSP	w/o	57.3	44.2	25.4
GrowSP*	w	57.4(+0.1)	45.9	27.1
Ours	w/o	77.0	57.8	43.8
Ours*	w	**78.7(+1.7)**	**59.2**	**46.7**

where the eigenvectors $\boldsymbol{U} \in \mathbb{R}^{S\times S}$ serve as the graph Fourier basis. Each eigenvector $\boldsymbol{u}_s \in \mathbb{R}^{S\times 1}$ represents a specific global structural pattern defined over all S superpoints.

Frequency-domain Pattern Grouping Directly using all S global patterns is computationally expensive and contains semantic redundancies. To obtain discriminative semantic priors, we group these patterns as follows:

- **Spectral Projection:** We project the spatial features $\boldsymbol{F}$ into the graph frequency domain to obtain frequency features $\boldsymbol{F}_{feq} \in \mathbb{R}^{S\times C}$:

$$\boldsymbol{F}_{feq} = \boldsymbol{U}^\top \boldsymbol{F}, \quad \boldsymbol{F}_{feq} = [\boldsymbol{f}_{feq,1}, \ldots, \boldsymbol{f}_{feq,s}, \ldots, \boldsymbol{f}_{feq,S}]^\top, \tag{9}$$

  where each row $\boldsymbol{f}_{feq,s} \in \mathbb{R}^{1\times C}$ represents the frequency response corresponding to the s-th global pattern $\boldsymbol{u}_s$.

- **K-means Clustering:** We apply K-means clustering to the frequency vectors $\{\boldsymbol{f}_{feq,1}, \ldots, \boldsymbol{f}_{feq,S}\}$ to partition the S original patterns into S' clusters.

- **Pattern Aggregation:** For each cluster, we average its assigned eigenvectors to form a refined global pattern $\boldsymbol{v}_{s'}$:

$$\boldsymbol{v}_{s'} = \frac{1}{W}\sum_{w=1}^{W} \boldsymbol{u}_w, \quad \boldsymbol{v}_{s'} \in \mathbb{R}^{S\times 1}, \tag{10}$$

  where W is the number of original patterns in the s'-th cluster.

The resulting matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{S'}] \in \mathbb{R}^{S\times S'}$ captures consolidated semantic priors by grouping patterns with similar spectral responses. This refined basis $\boldsymbol{V}$ is subsequently used for global branch.

D Additional Experiments and Quantitative Results

Controlled Analysis of External Model Priors. To investigate whether the performance improvement of LangTail primarily comes from the external foundation models rather than our proposed architectural design, we conduct a controlled experiment by augmenting the strong baselines, LogoSP and GrowSP, with the same external model priors used by our method, including VLM, SAM3 and CLIP features. Specifically, the corresponding baselines are denoted by LogoSP* and GrowSP*, where the superscript $*$ indicates that the same external model priors are incorporated. We report the results before and after introducing these priors in Table 6.

The results show that incorporating the external language-guided priors improves both baseline methods, demonstrating that the proposed language-guided module can provide useful information beyond the original representations of these methods. However, even after being equipped with the same external priors, LogoSP* and GrowSP* remain substantially below our method. This comparison suggests that the performance gains of LangTail arise from both the effectiveness of the external language-guided priors and the proposed architectural design, rather than from the external models alone.

Table 7: Controlled experiments on the effect of external foundation-model priors on ScanNet dataset.

Method	category-agnostic	mAcc(%)	mIoU(%)
CLIP2Scene	×	-	25.6
CNS	×	-	33.5
CUS3D	×	75.9	57.4
OpenScene-OpenSeg	×	70.7	47.5
OpenScene-LSeg	×	66.6	54.2
LangTail (ours)	✓	59.2	46.7

Table 8: Controlled experiments on the effect of external foundation-model priors on nuScenes dataset.

Method	category-agnostic	mAcc(%)	mIoU(%)
CLIP2Scene	×	-	20.8
CNS	×		26.8
OpenScene-OpenSeg	×	61.8	42.1
OpenScene-LSeg	×	42.7	36.7
LangTail (ours)	✓	35.1	29.0

Task Setting Clarification and Comparison with Label-Free Text-Prompted Segmentation Several recent works, including CLIP2Scene Chen et al. (2023b), CNS Chen et al. (2023a), CUS3D Yu et al. (2024), SAL Ošep et al. (2024), and OpenScene Peng et al. (2023), investigate label-free 3D segmentation using vision-language priors. We clarify that these methods and LangTail differ in their task protocols, particularly in whether the target semantic categories are provided as input.

Task protocol. Label-free text-prompted 3D segmentation methods take a 3D point cloud together with predefined semantic categories (e.g., text prompts) and predict the corresponding 3D masks. The absence of point-wise annotations in these methods therefore does not imply that the semantic categories are unknown: the target category vocabulary is provided as part of the input.

In contrast, our setting is *category-agnostic*. Given only 3D point clouds, the model must discover semantically meaningful groups and their corresponding 3D masks without predefined category names. We subsequently evaluate the discovered clusters against human-defined semantic classes using Hungarian matching. Thus, the key distinction is not whether external knowledge is used, but whether the target semantic categories are specified beforehand.

Relation to existing work. These label-free text-prompted methods are highly relevant to our work because they demonstrate the usefulness of vision-language knowledge for 3D segmentation. We therefore discuss them explicitly and report their published results in Table D and Table D. However, their text-conditioned protocol and our category-agnostic protocol address different levels of supervision. The former provides the semantic vocabulary through text prompts, whereas ours must discover the semantic categories jointly with the segmentation structure.

ScanNet. Tables 9 and Table 10 provide detailed results for each category on the validation and hidden test sets. Our method significantly outperforms all unsupervised baseline models, particularly in challenging subcategories such as *sinks*, *bathtubs*, and *refrigerators*.

S3DIS. The per-category results for each area evaluation are presented in Table 11 to Table 16 .Our method outperformed the unsupervised baseline method in all experiments.It has also achieved greatest performance in many specific categories

nuScenes. The table 17 shows the detailed results for each category. Our results outperform all baselines, demonstrating that our method is effective on outdoor datasets as well.

Table 9: Quantitative results on the validation split of ScanNet dataset. All 20 categories are evaluated.

Method	OA(%)	mAcc(%)	mIoU(%)	wall.	floor.	cab.	bed.	chair.	sofa.	table	door.	wind.	books.	pic.	counter.	desk.	curtain.	fridge.	shower.	toilet.	sink.	bathtub.	otherf.
RandCNN	11.9	8.4	3.2	9.3	10.0	3.5	2.5	6.8	2.0	4.8	5.1	3.9	3.1	1.7	0.8	2.3	2.8	1.0	0.2	0.2	0.1	0.6	3.6
van Kmeans	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	0.7	0.8	0.3	0.9	4.0
van Kmeans-S	10.2	9.8	3.4	8.9	10.3	3.4	3.2	5.5	3.4	4.2	3.4	5.2	3.1	**2.6**	0.7	2.8	4.2	0.6	0.8	0.7	0.2	1.0	4.1
van Kmeans-PFH	10.4	10.3	3.5	8.6	12.7	2.9	2.8	4.5	3.2	3.6	3.7	6.3	4.0	2.4	1.0	2.9	3.2	1.0	1.0	0.6	0.4	1.1	3.5
van Kmeans-S-PFH	12.2	9.3	3.6	11.3	12.3	2.9	2.4	5.4	2.8	4.2	3.8	5.8	3.8	2.3	1.2	2.4	2.9	0.9	1.4	0.6	0.1	1.1	4.1
IIC Ji et al. (2019)	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
IIC-S Ji et al. (2019)	18.3	6.7	3.4	18.3	16.0	2.6	2.3	4.4	2.0	5.4	3.2	2.9	3.3	0.7	0.4	1.4	1.6	0.7	0.1	0.3	0.1	0	2.6
IIC-PFH Ji et al. (2019)	25.4	6.3	3.4	29.6	14.9	1.1	1.0	5.6	0.8	3.6	3.0	1.6	1.3	0	0.3	1.0	0.4	0.4	0.2	0	0.1	0	3.2
IIC-S-PFH Ji et al. (2019)	18.9	6.3	3.0	18.0	15.9	3.4	0.9	7.1	0.6	0.8	4.3	1.6	3.5	0.4	0.1	0.3	0.3	0	0	0.1	0	0	2.7
PICIE Cho et al. (2021)	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	0.5	5.0
PICIE-S Cho et al. (2021)	35.6	13.7	8.1	38.4	53.9	4.3	2.7	10.2	6.3	14.1	5.2	4.0	6.0	0.2	1.3	2.1	1.5	0.2	0	2.6	3.1	1.3	4.3
PICIE-PFH Cho et al. (2021)	33.1	14.0	8.1	34.7	54.8	3.9	5.4	13.3	6.5	11.7	4.2	3.8	6.5	0.5	1.0	2.6	5.0	1.3	1.0	0.6	0	0.7	4.4
PICIE-S-PFH Cho et al. (2021)	23.6	15.1	7.4	18.1	39.1	5.4	4.9	13.4	6.9	20.3	5.8	4.5	7.7	1.2	3.0	5.8	4.7	0.6	1.2	0.4	0	1.1	4.5
PointDC Chen et al. (2023d)	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 Chen et al. (2023d)	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
GrowSP Zhang et al. (2023)	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
LogoSP Zhang et al. (2025)	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	0.0	46.3	0.0	42.3	29.6
GrowSP++ Zhang et al. (2026)	70.5	48.8	33.2	63.6	87.9	21.9	65.3	63.6	49.8	45.0	15.5	42.9	61.8	0.1	35.7	27.4	34.9	2.1	7.8	26.3	0	0	12.0
LangTail(Ours)	**78.7**	**59.2**	**46.7**	**69.7**	90.0	**42.1**	**67.9**	**69.8**	**52.8**	**51.1**	**37.9**	**44.2**	**64.1**	0.0	**43.5**	**40.2**	**40.0**	**3.5**	**40.1**	**67.7**	**20.9**	**52.4**	**35.5**

Table 10: Quantitative results on the online hidden split of ScanNet dataset. All 20 categories are evaluated.

Method		mIoU(%)	wall.	floor.	cab.	bed.	chair.	sofa.	table	door.	wind.	books.	pic.	counter.	desk.	curtain.	fridge.	shower.	toilet.	sink.	bathtub.	otherf.
Supervised Methods	PointNet++ Qi et al. (2017b)	33.9	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
	DGCNN Wang et al. (2019)	44.6	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
	PointCNN Li et al. (2018)	45.8	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
	SparseConv Graham et al. (2018)	72.5	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
Unsupervised Methods	GrowSP	26.9	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
	LogoSP Zhang et al. (2025)	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
	GrowSP++ Zhang et al. (2026)	32.3	32.3	66.4	86.5	14.7	58.9	**55.5**	**46.2**	**38.1**	26.2	47.3	49.9	0.0	29.0	29.0	33.6	0.0	0.0	38.9	0.0	11.4
	ours	**45.3**	**70.8**	**89.5**	36.0	63.1	**67.1**	55.2	39.8	34.7	46.9	46.8	0.0	28.6	**36.9**	**47.9**	**33.8**	**47.9**	**68.2**	16.7	**48.0**	**27.1**

E LIMITATIONS

While langTail achieves outperforming results across multiple benchmarks, several limitations remain.

Limited 2D coverage in outdoor scenes. langTail relies on multi-view RGB images to distill semantic priors from 2D to 3D via DINOv2. However, outdoor LiDAR datasets often provide sparse or temporally misaligned image coverage with narrow fields of view, resulting in insufficient 2D-3D correspondences for effective distillation.

Sensitivity to pseudo-label quality. The entity branch leverages a pre-generated vocabulary to guide semantic discovery and pseudo-label assignment. Its stability is therefore contingent on the coverage, precision, and scene-relevance of this vocabulary. If the generation process yields redundant, ambiguous, or domain-mismatched terms, cross-modal alignment may propagate noise into the clustering pipeline.

Table 11: Quantitative results of our method and baselines on the Area-1 of S3DIS dataset.

Method		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised Methods	PointNet Qi et al. (2017a)	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++ Qi et al. (2017b)	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 Graham et al. (2018)	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 Methods	RandCNN	24.9	20.3	10.9	27.8	29.0	16.6	10.7	4.7	14.8	8.1	4.1	3.7	0.2	6.4	5.1
	van 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
	van Kmeans-S	21.8	25.6	10.8	18.1	20.4	9.5	14.5	1.0	17.2	16.2	9.6	6.8	0.6	8.2	7.6
	van Kmeans-PFH	26.4	25.7	12.6	34.0	27.4	11.6	17.0	5.2	15.4	9.7	10.8	5.8	1.6	8.1	4.3
	van Kmeans-S-PFH	25.1	23.0	10.8	26.1	18.5	15.0	4.6	2.6	20.5	15.9	10.4	6.0	0.6	5.9	3.7
	IIC Ji et al. (2019)	29.2	14.3	8.0	17.0	31.4	25.6	4.3	11.1	0.0	2.6	1.4	0.7	0.0	0.2	1.4
	IIC-S Ji et al. (2019)	31.3	21.1	8.2	35.5	11.6	13.9	1.3	0.2	13.0	15.6	5.1	0.6	0.3	0.3	7.6
	IIC-PFH Ji et al. (2019)	27.3	12.0	6.1	17.1	17.0	23.4	4.4	0.3	0.8	4.0	0.4	0.4	0.0	4.9	0.1
	IIC-S-PFH Ji et al. (2019)	24.1	13.2	7.2	21.7	16.8	18.0	5.0	3.6	2.9	3.7	4.4	0.7	0.0	4.7	4.5
	PICIE Cho et al. (2021)	45.7	28.3	19.4	77.2	63.1	24.5	15.8	3.3	4.4	9.6	10.2	14.7	0.0	9.9	0.0
	PICIE-S Cho et al. (2021)	48.9	30.3	21.8	77.7	58.1	39.0	13.6	2.5	2.8	11.8	20.0	17.1	1.8	17.1	0.6
	PICIE-PFH Cho et al. (2021)	51.9	38.4	27.1	59.8	78.7	37.5	6.8	9.5	4.9	26.4	50.2	29.5	0.0	22.8	0.1
	PICIE-S-PFH Cho et al. (2021)	47.9	38.9	27.2	71.5	66.3	23.9	20.4	8.8	6.9	12.9	39.3	32.2	0.2	42.0	2.4
	PointDC Chen et al. (2023d)	58.0	41.5	28.8	88.7	89.5	31.9	1.5	7.1	17.6	12.4	46.4	17.0	0.0	32.7	0.0
	PointDC-DINOv2 Chen et al. (2023d)	73.8	55.4	44.0	85.7	**93.7**	58.9	10.1	**20.2**	0.0	45.0	70.1	61.5	44.0	39.4	0.0
	GrowSP Zhang et al. (2023)	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
	LogoSP Zhang et al. (2025)	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	0.0
	GrowSP++ Zhang et al. (2026)	72.9	62.9	47.5	78.8	92.9	53.2	0.0	1.1	74.4	33.9	66.8	71.4	54.8	43.1	0.1
	LangTail(ours)	**84.3**	**74.6**	**63.6**	92.8	90.9	**68.7**	**45.0**	1.8	**76.3**	**66.1**	**76.3**	**79.4**	**67.5**	**62.5**	**35.6**

Table 12: Quantitative results of our method and baselines on the Area-2 of S3DIS dataset.

Method		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised Methods	PointNet Qi et al. (2017a)	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++ Qi et al. (2017b)	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 Graham et al. (2018)	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	8.15
Unsupervised Methods	RandCNN	20.4	17.0	7.6	16.8	19.2	12.8	1.6	1.1	1.5	9.4	2.9	16.1	0.3	6.4	**13.0**
	van 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
	van Kmeans-S	17.3	17.6	6.3	14.7	15.0	12.1	3.0	0.9	0.0	7.0	3.6	11.4	1.9	4.3	2.0
	van Kmeans-PFH	21.8	22.3	9.0	31.0	21.1	10.1	3.0	2.6	3.2	10.7	5.3	9.8	4.0	7.2	0.6
	van Kmeans-S-PFH	21.5	18.0	8.0	23.7	21.1	14.1	3.8	1.8	1.8	5.4	4.3	10.2	2.5	6.5	0.9
	IIC Ji et al. (2019)	41.6	16.8	10.6	33.0	43.7	27.6	1.7	0.0	0.0	5.6	0.1	13.0	0.0	2.8	0.0
	IIC-S Ji et al. (2019)	30.3	11.9	6.2	29.3	7.6	22.6	0.5	1.5	0.0	3.5	0.3	2.2	1.8	3.4	1.1
	IIC-PFH Ji et al. (2019)	27.3	10.5	5.8	16.0	13.9	24.9	0.3	0.1	0.0	4.0	0.6	8.3	0.1	0.8	0.1
	IIC-S-PFH Ji et al. (2019)	26.8	12.8	6.8	30.5	12.5	16.5	1.1	0.4	1.1	7.0	1.1	5.7	0.4	4.0	1.2
	PICIE Cho et al. (2021)	48.3	27.2	17.4	72.4	44.2	39.6	6.2	1.7	0.5	7.7	4.1	20.1	0.0	7.7	3.6
	PICIE-S Cho et al. (2021)	55.5	30.4	21.5	69.1	68.7	46.8	3.1	1.1	0.4	11.6	6.6	38.2	3.0	7.7	1.8
	PICIE-PFH Cho et al. (2021)	55.8	37.2	25.7	51.7	82.6	41.8	2.5	5.3	1.0	13.7	40.6	55.0	0.0	12.5	2.1
	PICIE-S-PFH Cho et al. (2021)	55.4	39.2	26.2	65.8	77.7	29.3	4.6	2.7	0.5	12.4	26.7	62.2	1.0	28.0	3.7
	PointDC Chen et al. (2023d)	48.3	34.8	22.5	66.7	50.2	26.1	1.3	0.4	0.0	15.6	29.8	56.3	0.3	17.6	5.6
	PointDC-DINOv2 Chen et al. (2023d)	77.1	50.3	38.1	90.8	**92.0**	57.3	10.2	0.6	34.2	20.3	46.4	**85.6**	0.0	19.1	0.9
	GrowSP Zhang et al. (2023)	79.0	51.8	39.1	85.7	88.2	67.0	**12.0**	**24.8**	0.0	24.2	51.2	77.1	4.1	24.5	0.2
	LogoSP Zhang et al. (2025)	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
	GrowSP++ Zhang et al. (2026)	79.5	54.5	43.8	90.9	91.2	66.7	8.9	0.4	**67.5**	43.7	57.0	54.4	0.3	43.8	0.8
	LangTail(ours)	**88.9**	**62.9**	**53.4**	**92.4**	91.7	**78.8**	1.8	23.5	67.2	**63.4**	**58.4**	83.7	**18.5**	**55.2**	6.9

Table 13: Quantitative results of our method and baselines on the Area-3 of S3DIS dataset.

Method		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised Methods	PointNet Qi et al. (2017a)	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++ Qi et al. (2017b)	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 Graham et al. (2018)	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	52.2
Unsupervised Methods	RandCNN	25.3	21.7	10.8	25.5	32.6	17.2	4.9	3.3	10.6	9.1	4.2	2.9	0.9	9.3	8.8
	van 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
	van Kmeans-S	21.2	21.9	9.3	19.8	18.8	14.4	5.0	1.6	1.5	14.4	7.4	6.6	4.2	10.3	7.7
	van Kmeans-PFH	24.7	27.3	11.6	33.6	23.0	11.0	8.9	3.5	10.8	12.0	8.4	6.9	3.7	12.6	4.9
	van Kmeans-S-PFH	23.3	21.7	9.7	24.7	20.1	14.1	5.4	2.3	5.2	12.3	5.1	7.3	3.1	9.3	8.0
	IIC Ji et al. (2019)	32.1	15.4	8.4	20.5	25.5	31.4	1.0	6.9	0.2	3.2	1.6	0.3	0.0	10.6	0.0
	IIC-S Ji et al. (2019)	30.3	14.8	7.1	33.9	10.1	16.1	2.0	0.4	5.1	8.8	0.6	0.3	2.7	2.1	5.5
	IIC-PFH Ji et al. (2019)	34.5	11.8	6.3	22.6	13.0	32.4	0.1	0.0	0.1	1.9	0.3	0.3	1.0	2.3	1.9
	IIC-S-PFH Ji et al. (2019)	25.5	12.8	6.6	26.1	8.7	18.8	1.3	2.9	0.8	5.3	1.6	4.8	1.7	5.3	2.6
	PICIE Cho et al. (2021)	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
	PICIE-S Cho et al. (2021)	52.8	32.3	22.4	79.4	80.0	38.6	9.1	2.3	3.9	10.7	15.9	11.0	0.0	17.6	0.6
	PICIE-PFH Cho et al. (2021)	57.8	36.7	26.4	60.5	78.8	53.9	4.1	4.8	0.3	16.0	50.0	25.3	1.0	26.4	0.6
	PICIE-S-PFH Cho et al. (2021)	51.8	43.3	29.3	85.0	72.6	23.7	14.2	4.0	5.7	12.7	48.5	32.2	2.8	48.3	2.1
	PointDC Chen et al. (2023d)	56.2	40.7	27.2	75.3	91.3	29.7	1.2	2.2	0.0	11.4	37.9	20.7	9.2	38.4	8.9
	PointDC-DINOv2 Chen et al. (2023d)	70.5	52.4	39.3	86.2	93.4	48.8	0.0	14.5	38.6	28.2	68.7	55.0	0.0	37.7	0.0
	GrowSP Zhang et al. (2023)	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	0.0
	LogoSP Zhang et al. (2025)	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	0.0
	GrowSP++ Zhang et al. (2026)	78.7	65.6	50.2	92.3	94.0	59.3	30.6	15.5	68.5	53.1	60.7	56.2	0.6	71.2	0.0
	LangTail(ours)	**87.1**	**72.5**	**62.6**	91.1	91.5	**74.9**	13.5	0.0	**79.0**	**68.4**	**72.8**	**71.1**	**67.4**	**77.4**	**43.8**

Table 14: Quantitative results of our method and baselines on the Area-4 of S3DIS dataset.

Method		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised Methods	PointNet Qi et al. (2017a)	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++ Qi et al. (2017b)	74.8	66.4	47.7	85.5	96.1	68.9	4.4	23.8	27.0	50.5	44.9	54.0	35.6	38.4	43.8
	SparseConv Graham et al. (2018)	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 Methods	RandCNN	22.6	18.3	9.0	22.0	28.2	16.4	0.4	4.2	5.6	11.9	4.5	4.3	1.4	6.4	2.3
	van 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
	van Kmeans-S	17.3	19.6	7.4	16.7	14.8	10.2	0.9	4.4	5.1	10.5	6.1	7.5	2.8	9.8	0.0
	van Kmeans-PFH	21.7	22.3	9.3	28.9	21.3	12.6	1.2	5.9	4.4	11.1	6.6	8.6	2.6	7.4	0.7
	van Kmeans-S-PFH	23.1	18.3	8.9	24.4	26.2	14.5	1.7	5.0	3.2	11.7	3.8	6.1	0.4	9.0	1.1
	IIC Ji et al. (2019)	33.0	13.5	8.2	14.9	25.9	35.1	0.0	1.1	1.4	3.7	4.1	0.9	0.0	10.8	0.0
	IIC-S Ji et al. (2019)	26.7	13.5	6.1	27.6	4.9	15.3	0.1	1.0	5.3	12.0	1.5	2.3	1.1	1.6	0.0
	IIC-PFH Ji et al. (2019)	31.0	11.4	6.3	21.0	10.9	30.0	0.1	1.3	3.0	2.3	0.8	1.3	0.0	4.6	0.1
	IIC-S-PFH Ji et al. (2019)	27.3	13.2	6.9	29.4	9.9	20.0	0.1	2.6	3.1	4.0	3.7	2.5	0.6	6.3	0.8
	PICIE Cho et al. (2021)	43.2	29.4	17.8	62.2	72.7	22.6	2.5	3.4	3.5	8.8	4.1	17.4	0.0	15.5	0.7
	PICIE-S Cho et al. (2021)	50.0	32.2	22.3	72.7	77.7	37.4	1.6	2.4	3.3	11.3	25.8	19.7	3.5	12.3	0.0
	PICIE-PFH Cho et al. (2021)	58.6	42.6	28.5	69.5	83.6	48.3	2.6	9.4	5.8	16.4	44.1	36.3	1.6	23.9	0.1
	PICIE-S-PFH Cho et al. (2021)	47.1	44.8	27.6	67.2	67.8	25.9	2.6	14.5	5.2	10.2	47.9	44.1	0.9	44.2	0.4
	PointDC Chen et al. (2023d)	54.0	35.3	25.2	87.9	86.8	24.7	0.0	4.2	12.2	18.7	32.6	17.6	1.9	15.6	0.0
	PointDC-DINOv2 Chen et al. (2023d)	73.0	46.5	39.8	89.4	91.9	58.2	0.0	0.7	25.3	19.0	52.4	53.8	0.7	59.0	0.0
	GrowSP Zhang et al. (2023)	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	0.0
	LogoSP Zhang et al. (2025)	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**
	GrowSP++ Zhang et al. (2026)	77.3	52.6	42.8	91.0	92.3	65.0	0.0	6.4	42.0	47.1	52.7	60.7	0.5	55.2	0.0
	LangTail(ours)	**87.6**	**70.6**	**59.5**	90.3	93.3	**76.9**	0.0	**39.4**	**50.7**	**63.6**	**72.5**	**74.9**	55.8	**76.1**	21.2

Table 15: Quantitative results of our method and baselines on the Area-5 of S3DIS dataset.

Method		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised Methods	PointNet Qi et al. (2017a)	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++ Qi et al. (2017b)	77.5	62.6	50.1	83.1	97.2	66.4	0.0	8.1	55.6	15.2	60.4	64.5	36.6	58.3	55.7
	SparseConv Graham et al. (2018)	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	30.39
Unsupervised Methods	RandCNN	23.3	17.3	9.2	25.3	24.5	17.4	0.0	2.3	12.5	6.5	5.7	3.0	0.3	10.1	2.2
	van 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	0.0	13.6	2.3
	van Kmeans-S	21.9	22.9	9.0	19.3	18.1	17.0	0.2	2.1	11.8	4.5	8.9	6.6	0.2	14.0	4.8
	van Kmeans-PFH	23.2	23.6	10.2	32.0	20.5	10.3	0.1	3.6	15.2	7.1	9.9	6.2	0.7	12.4	4.9
	van Kmeans-S-PFH	22.8	20.6	9.2	25.2	26.5	12.7	0.4	2.0	8.7	8.1	5.8	5.7	0.0	12.7	3.3
	IIC Ji et al. (2019)	28.5	12.5	6.4	6.1	19.8	27.9	0.0	2.1	0.1	3.4	7.9	0.4	0.0	8.6	0.0
	IIC-S Ji et al. (2019)	29.2	13.0	6.8	28.9	12.3	18.7	0.0	0.1	3.6	1.3	3.8	0.6	0.0	8.1	3.8
	IIC-PFH Ji et al. (2019)	28.6	16.8	7.9	23.7	24.9	17.7	5.9	1.4	12.6	0.2	5.3	0.6	0.0	2.4	0.0
	IIC-S-PFH Ji et al. (2019)	31.2	16.3	9.1	43.2	23.6	14.9	0.0	1.6	3.9	2.4	3.6	1.5	0.8	9.5	4.5
	PICIE Cho et al. (2021)	61.6	25.8	17.9	65.7	61.4	58.4	0.0	0.3	2.2	1.7	12.1	0.0	0.0	12.4	0.0
	PICIE-S Cho et al. (2021)	49.6	28.9	20.0	64.2	75.1	42.4	0.1	1.2	4.6	7.4	18.7	9.2	1.0	16.0	0.4
	PICIE-PFH Cho et al. (2021)	54.0	36.8	24.4	58.4	68.6	49.9	0.1	7.6	5.3	14.2	45.1	16.3	0.1	27.1	0.6
	PICIE-S-PFH Cho et al. (2021)	48.4	40.4	25.2	59.6	72.5	26.0	0.2	8.5	5.9	8.7	46.0	26.9	0.4	46.8	0.3
	PointDC Chen et al. (2023d)	55.5	35.1	23.9	84.4	84.3	30.2	0.0	1.8	12.2	7.1	24.6	6.9	5.4	29.7	0.7
	PointDC-DINOv2 Chen et al. (2023d)	75.7	48.7	40.2	87.7	89.5	59.2	0.0	0.8	25.8	26.3	62.0	68.3	1.5	61.0	0.5
	GrowSP Zhang et al. (2023)	78.4	57.2	44.5	90.5	90.1	66.7	0.0	**14.8**	27.6	**45.6**	59.4	71.9	10.7	56.0	0.2
	LogoSP Zhang et al. (2025)	82.8	55.9	46.5	**92.9**	**95.4**	73.2	0.0	3.3	57.8	35.9	55.5	**74.6**	1.9	67.3	0.3
	GrowSP++ Zhang et al. (2026)	78.7	60.1	46.6	87.3	92.8	53.5	0.0	5.3	54.2	24.4	56.8	63.4	52.1	69.0	0.0
	LangTail(ours)	**86.7**	**69.6**	**59.5**	92.5	90.5	**73.8**	**8.2**	8.7	**58.0**	38.0	**73.3**	71.0	**68.7**	**76.7**	**54.9**

Table 16: Quantitative results of our method and baselines on the Area-6 of S3DIS dataset.

Method		OA(%)	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board.
Supervised Methods	PointNet Qi et al. (2017a)	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++ Qi et al. (2017b)	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 Graham et al. (2018)	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 Methods	RandCNN	22.1	16.0	8.5	17.6	24.2	19.2	0.0	1.7	12.2	7.6	6.2	2.6	0.2	8.9	1.7
	van 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
	van Kmeans-S	20.6	26.3	10.2	16.6	17.8	9.0	13.4	0.6	15.0	17.0	10.7	8.3	1.8	7.4	5.1
	van Kmeans-PFH	25.8	27.2	12.8	32.0	25.1	12.2	15.0	3.6	19.7	14.9	9.4	10.9	2.1	5.3	3.5
	van Kmeans-S-PFH	23.9	23.0	10.5	22.2	20.8	15.2	14.9	0.3	9.5	13.6	10.9	7.4	1.6	5.0	4.9
	IIC Ji et al. (2019)	32.5	15.9	9.2	21.9	33.8	29.1	3.1	15.2	0.0	2.7	0.7	0.0	0.0	1.5	1.8
	IIC-S Ji et al. (2019)	27.0	16.5	7.6	28.9	8.5	12.7	0.5	0.2	0.0	14.1	10.4	0.6	0.0	1.0	3.8
	IIC-PFH Ji et al. (2019)	28.4	16.7	7.8	23.6	24.5	17.4	5.8	1.5	12.6	0.2	4.8	0.6	0.0	2.4	0.0
	IIC-S-PFH Ji et al. (2019)	22.9	13.1	6.7	28.8	9.0	12.6	0.6	4.8	2.1	6.6	4.4	1.2	0.3	5.4	4.9
	PICIE Cho et al. (2021)	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	0.0
	PICIE-S Cho et al. (2021)	47.4	30.7	21.8	65.5	67.5	37.2	16.6	2.0	4.8	10.6	23.7	23.4	23.4	1.7	0.0
	PICIE-PFH Cho et al. (2021)	51.8	41.3	27.9	63.1	56.5	39.0	11.4	10.0	5.3	19.1	63.8	50.4	0.5	37.2	0.9
	PICIE-S-PFH Cho et al. (2021)	44.0	36.1	24.7	58.2	60.4	24.5	17.9	10.1	8.1	12.7	44.0	5.5	0.5	54.3	0.3
	PointDC Chen et al. (2023d)	62.4	38.7	28.6	85.8	85.6	43.8	5.0	16.5	8.8	10.7	41.7	12.5	0.0	33.0	0.0
	PointDC-DINOv2 Chen et al. (2023d)	76.4	55.5	46.4	90.3	91.4	61.7	0.0	19.7	63.1	33.6	67.9	65.7	1.4	62.5	0.0
	GrowSP Zhang et al. (2023)	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
	LogoSP Zhang et al. (2025)	77.9	62.9	50.6	**94.6**	**92.7**	64.0	25.7	23.9	65.9	38.8	68.9	72.2	2.5	68.4	0.0
	GrowSP++ Zhang et al. (2026)	78.4	63.5	51.5	86.8	92.6	63.9	36.4	21.2	**73.4**	39.8	66.6	71.8	1.8	64.7	0.0
	LangTail(ours)	**85.1**	**82.4**	**67.2**	88.7	88.3	**69.0**	47.7	**45.5**	68.9	**73.0**	**75.5**	**72.4**	**63.4**	**71.5**	**42.4**

Table 17: Per-category quantitative results on the validation split of nuScenes dataset.

Method	OA(%)	mAcc(%)	mIoU(%)	barrier	bicycle	bus	car	cons. veh.	motor.	ped.	cone	trailer	truck	driv. surf.	other flat	sidewalk	terrain	manmade	veget.
GrowSP Zhang et al. (2023)	39.2	17.5	10.2	7.5	0.0	0.4	42.9	0.1	0.0	0.6	0.0	0.7	1.4	48.4	0.8	6.5	13.1	21.4	19.7
PointDC Chen et al. (2023d)	56.8	29.4	17.7	11.6	0.0	0.5	63.1	0.3	0.0	4.4	0.0	1.2	26.4	70.1	0.1	7.1	19.3	21.1	58.1
PointDC-DINOv2 Chen et al. (2023d)	51.8	28.6	18.2	**17.0**	0.0	0.2	58.4	0.2	0.0	1.5	0.0	1.6	43.3	**71.8**	0.0	**8.3**	19.5	17.6	51.8
LogoSP Zhang et al. (2025)	54.8	29.2	20.1	16.6	0.0	0.7	70.2	0.2	0.2	**33.6**	0.0	0.3	38.4	59.4	0.4	8.0	10.7	33.0	49.3
GrowSP++ Zhang et al. (2026)	55.7	30.4	19.7	16.5	0.0	33.8	73.0	**0.4**	0.0	0.8	0.0	1.0	44.7	63.5	**1.1**	3.6	12.7	20.6	49.9
Ours	**73.9**	**35.1**	**29.0**	6.0	0.0	**50.2**	**74.0**	0.0	**5.0**	26.3	0.0	**13.7**	**51.6**	68.4	0.0	6.1	**23.7**	**71.7**	**66.9**