No Object Is an Island: Enhancing 3D Semantic Segmentation Generalization with Diffusion Models

Fan Li [1] **Xuan Wang** [1] **Xuanbin Wang**[1] **Zhaoxiang Zhang**[1] **Yuelei Xu**[1*]
[1]Northwestern Polytechnical University
lifan.messages@gmail.com

Abstract

Enhancing the cross-domain generalization of 3D semantic segmentation is a pivotal task in computer vision that has recently gained increasing attention. Most existing methods, whether using consistency regularization or cross-modal feature fusion, focus solely on individual objects while overlooking implicit semantic dependencies among them, resulting in the loss of useful semantic information. Inspired by the diffusion model's ability to flexibly compose diverse objects into high-quality images across varying domains, we seek to harness its capacity for capturing underlying contextual distributions and spatial arrangements among objects to address the challenging task of cross-domain 3D semantic segmentation. In this paper, we propose a novel cross-modal learning framework based on diffusion models to enhance the generalization of 3D semantic segmentation, named XDiff3D. XDiff3D comprises three key ingredients: (1) constructing object agent queries from diffusion features to aggregate instance semantic information; (2) decoupling fine-grained local details from object agent queries to prevent interference with 3D semantic representation; (3) leveraging object agent queries as an interface to enhance the modeling of object semantic dependencies in 3D representations. Extensive experiments validate the effectiveness of our method, achieving state-of-the-art performance across multiple benchmarks in different task settings. Code is available at `https://github.com/FanLiHub/XDiff3D`.

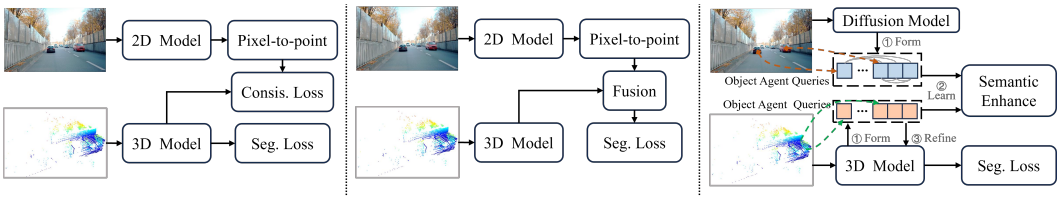
1 Introduction

3D semantic segmentation, a fundamental task in computer vision with widespread applications in autonomous driving, robotics, and augmented reality, has made significant advancements in recent years [39, 40, 32, 48, 17, 10]. Despite these advancements, it still experiences severe performance degradation when models trained on the source domain are applied to unseen target domains due to the existence of a domain gap. This has sparked growing interest in Domain Generalized 3D Semantic Segmentation (DG3SS) [28, 62, 24, 55, 68], aiming to learn domain-invariant features that enable models to perform well on a variety of unseen target domains with similar semantic distribution.

Current approaches can be broadly categorized into uni-modal methods based solely on point clouds and cross-modal methods that integrate both point cloud and image data, as shown in Figure 1. The former focuses on reducing the domain gap between the source and target domains through domain augmentation [24, 42, 44] or domain mixing [43, 25], but its performance remains limited. The latter exploits image features paired with point clouds to perform consistency regularization [37, 19, 59, 55] or cross-modal feature fusion [54, 28], substantially improving cross-domain generalization and delivering impressive results. However, most existing approaches *treat objects as isolated islands*, focusing solely on domain-invariant features of individual objects and overlooking the implicit

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Consistency Regularization (b) Cross-modal Fusion (c) Our Method

Figure 1: Existing methods mainly adopt two paradigms: (a) consistency regularization, which enforces alignment across modalities, and (b) cross-modal feature fusion, which integrates multi-modal information. However, both paradigms focus primarily on individual domain-invariant features, overlooking rich semantic dependencies among objects. In contrast, our method (c) leverages object agent queries as an interface to incorporate instance semantic dependencies from diffusion priors into the 3D semantic space, enhancing cross-domain generalization of 3D semantic segmentation.

semantic dependencies among them such as spatial arrangements and contextual relations, which leads to the loss of informative cues and ultimately leads to suboptimal performance.

Recently, diffusion models have provided a new perspective for improving semantic segmentation generalization with remarkable capabilities in capturing underlying semantic relations among objects to synthesize high-quality samples across diverse domains [18, 12, 41]. Building on this insight, several studies [38, 22, 35, 52] have leveraged diffusion priors to better model semantic dependencies among objects, thereby enhancing the generalization of 2D semantic segmentation. Given the remarkable progress of diffusion models in the 2D visual domain, a natural curiosity has been raised: **How can the prior knowledge encoded in diffusion models be leveraged to improve the generalization of 3D semantic segmentation?**

A naive solution is to treat diffusion models as general feature extractors and utilize their features as supervisory signals or auxiliary information by pixel-to-point matching for 3D semantic segmentation. However, this comes with two new challenges: (1) *Strict pixel-to-point matching leads to significant loss of image features due to point cloud sparsity* [1, 37], *further aggravated by mainstream diffusion models operating in compressed latent spaces rather than pixel spaces, hindering effective cross-modal feature association.* (2) *Diffusion models are inherently designed for generative tasks, and their feature spaces not only model high-level instance semantics but also retain fine-grained local visual details (e.g., slogans and scrawl) that may introduce noise and disturb 3D feature learning.*

To address the above challenges, we propose XDiff3D, a cross-modal learning framework via diffusion models for cross-domain 3D semantic segmentation, which leverages the semantic dependencies among objects embedded in pretrained diffusion priors to enrich the learned representations of point clouds. Specifically, for **the first challenge**, we introduce learnable queries that interact with diffusion features to aggregate rich semantic dependencies among objects, forming object agent queries that subsequently refine the 3D feature representations. To tackle **the second challenge**, we introduce a dual-query learning scheme that enforces semantic consistency between the principal components of two sets of queries, effectively suppressing interference from potentially distracting visual details. Ultimately, these object agent queries serve as an interface to infuse 2D semantic priors into the 3D representation space, enabling the refinement of 3D semantic features in the decoder for more generalizable representations. Extensive experiments across multiple benchmark settings demonstrate that XDiff3D consistently delivers state-of-the-art performance, surpassing existing baselines by a significant margin. Moreover, comprehensive ablation studies and analyses further validate the effectiveness of its core components. The contributions are summarized as follows:

- To the best of our knowledge, XDiff3D is the first diffusion-based cross-modal learning framework that constructs object agent queries as an interface to infuse 2D diffusion priors into the 3D representation space, thereby enhancing the cross-domain generalization of 3D semantic segmentation.
- We propose a dual-queries learning scheme that suppresses potentially intricate visual details embedded in diffusion priors to prevent interference with the learning of robust 3D feature representations.

- XDiff3D is a concise and general framework, consistently outperforming a wide variety of baselines and achieving state-of-the-art performance across multiple benchmarks for cross-domain 3D semantic segmentation.

2 Related Works

2.1 Diffusion Models

Diffusion models [18, 46, 12, 41] have demonstrated impressive capabilities in image generation and are increasingly being explored for visual perception tasks such as semantic segmentation [60, 53, 49], object detection [7, 20], and depth estimation [23]. Recent studies have begun to exploit diffusion models for improving domain generalization in 2D segmentation [38, 35, 4]. For example, DGInStyle [21] introduces a controllable framework that generates diverse, task-specific images from diffusion priors. Niemeijer et al. [35] use text-guided diffusion to synthesize pseudo-target domains for better coverage of target domain variations. Despite the progress in 2D tasks, the potential of diffusion priors to enhance generalization in 3D semantic segmentation remains underexplored. This work addresses this gap by leveraging instance-level semantic priors from diffusion models to enrich 3D semantic representations and improve their generalization with respect to domain shifts.

2.2 Domain Generalized 3D Semantic Segmentation

Domain Generalized 3D Semantic Segmentation (DG3SS) aims to learn a model solely from source domain data that can perform well across diverse unseen target domains, and has recently garnered significant attention in the research community [5, 62, 58, 55, 68]. SemanticSTF [58] introduces a large-scale benchmark for semantic segmentation of LiDAR point clouds in adverse weather, enabling comprehensive evaluation of domain adaptive and generalizable 3D segmentation methods under all-weather conditions. MM2D3D [5] injects depth cues into the 2D branch and RGB information into the 3D branch to improve modality complementarity and robustness to domain shift. 2DPASS [62] introduces a fusion-then-distillation strategy to transfer rich semantic and structural information from 2D images to 3D point clouds without requiring strictly paired data. UniDSeg [55] leverages Visual Foundation Models by introducing learnable prompts within a cross-modal framework to bridge the 2D-3D domain gap and enhance generalization in cross-domain 3D semantic segmentation. Despite these advancements, existing methods primarily focus on learning domain-invariant features for individual objects, while overlooking the latent semantic dependencies among objects. In contrast, this paper exploits the contextual relationships and spatial organization among objects to move beyond isolated instance modeling, enabling the model to grasp a unified semantic distribution underlying diverse scenes, thereby enhancing cross-domain generalization.

2.3 Learnable Queries Design

Recently, various approaches [36, 30, 66, 51] have adopted learnable query-based frameworks inspired by DETR [6]. MaskFormer [8] and Mask2Former [9] unify semantic and instance segmentation using object queries. Tqdm [36] introduces domain-invariant textual queries for domain generalized semantic segmentation. kMaX-DeepLab [64] treats pixel–query interaction as k-means-style clustering, simplifying cross-attention for better segmentation. While learnable queries have proven effective in various perception tasks, how to leverage them to enhance generalization in 3D semantic segmentation remains an open question. In this work, we derive object agent queries from diffusion features and utilize them as a cross-modal interface to enhance 3D semantic representations.

3 Method

3.1 Preliminary

Stable Diffusion. Stable diffusion models comprise two complementary stochastic processes: a diffusion process and a reverse process. During the diffusion process, random noise is progressively added to the data via a Markov chain:

$$q\left(z_t \mid z_0\right) := \mathcal{N}\left(z_t \mid \sqrt{\bar{\alpha}_t} z_0, \left(1-\bar{\alpha}_t\right) I\right), \quad z_0 = \mathcal{E}(x) \tag{1}$$

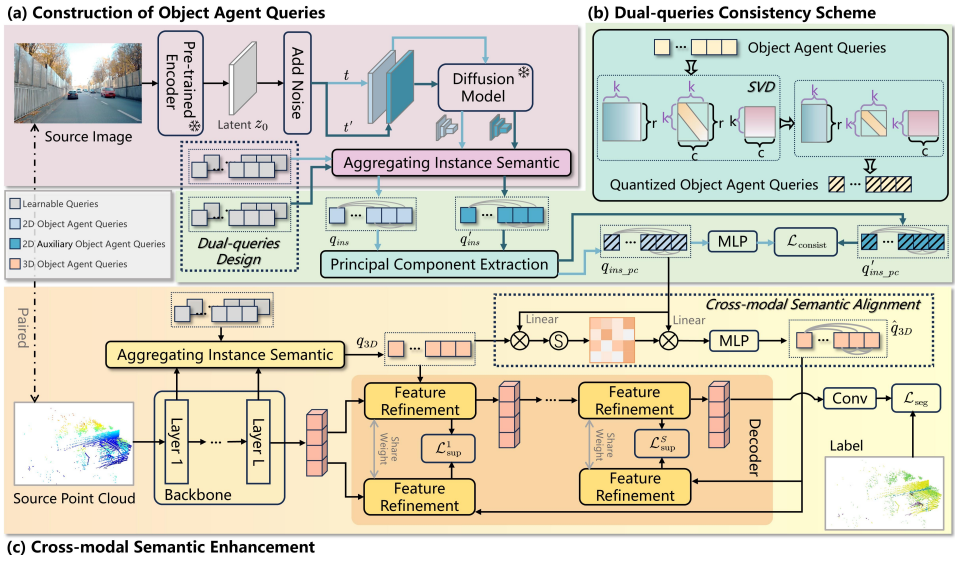


Figure 2: A brief illustration of our proposed framework. First, we aggregate semantic information of objects within the scene from diffusion features to form the object agent queries. Next, we propose a dual-query mechanism to eliminate local visual details from the object agent queries, preventing interference with 3D semantic representation. Finally, the optimized object agent queries are used as an interface to infuse inter-object semantic dependencies into the 3D representation, guiding the model to learn domain-invariant features.

where $\mathcal{E}$ denotes a pretrained VAE encoder that maps an input image $x \in \mathbb{R}^{H\times W\times 3}$ into a latent representation z_0. The hyperparameters $\bar{\alpha}_t$ represent a pre-defined noise schedule, where a larger t corresponds to larger noise weights. The reverse process gradually reconstructs clean data from noisy samples using a noise predictor $\epsilon_\theta(\cdot)$. Each step of this reverse process can be formulated as:

$$p_\theta\left(z_{t-1} \mid z_t\right) := \mathcal{N}\left(z_{t-1} \mid \mu_\theta\left(z_t, t\right), \Sigma_\theta\left(z_t, t\right)\right) \tag{2}$$

where μ_θ is the mean predicted by $\epsilon_\theta(\cdot)$, and Σ_θ is typically set to a predefined covariance value.

Problem Definition. The goal of DG3SS is to train a segmentation model solely on a labeled source domain $\mathcal{S}$, enabling it to generalize to unseen target domains $\mathcal{T}$. For the source domain, it is assumed that paired images and point clouds are available. In general, the segmentation model $\varphi = g \circ \upsilon$ consists of a backbone g for feature extraction and a decoder υ for producing semantic predictions.

3.2 Overview

With diffusion models successfully enhancing segmentation generalization in the 2D vision domain drawing on strong instance semantic priors, a natural question arises: How can the prior knowledge encoded in diffusion models be leveraged to bolster the generalization of 3D semantic segmentation? As illustrated in Figure 2, we propose XDiff3D, a cross-modal learning framework for cross-domain 3D semantic segmentation based on diffusion models, which consists of three key ingredients: (1) aggregating object semantic information from the prior knowledge encoded in diffusion features to generate object agent queries (Section 3.3), (2) suppressing potential intricate visual details embedded within the object agent queries (Section 3.4), and (3) leveraging the optimized object agent queries as an interface to perform cross-modal semantic enhancement (Section 3.5).

3.3 Construction of Object Agent Queries

Given that diffusion models operate in compressed latent spaces, traditional calibration-based hard associations between LiDAR points and image pixels become unreliable. To fully harness the rich instance semantic priors captured by diffusion features, we construct object agent queries from the diffusion space and employ them as an interface for infusing semantic knowledge into the 3D feature space, with the goal of enhancing the generalization capability of 3D semantic segmentation.

Diffusion Feature Extraction. Before that, we need to extract diffusion features. To do this, we begin by encoding the source image x into a latent code $z_0 = \mathcal{E}(x)$ via a pretrained encoder $\mathcal{E}$. Subsequently, we introduce noise to z_0 according to Equation 1, yielding the noisy latent code z_t. Finally, we apply the noise predictor $\epsilon_\theta(\cdot)$ to perform the denoising process to get diffusion features:

$$f_{sd}^{(i)} = \epsilon_\theta^i\left(z_t, t, \mathcal{T}(e)\right), \text{with } 1 \leq i \leq M \tag{3}$$

where $\mathcal{T}$ denotes the text encoder, $f_{sd}^{(i)}$ represents the i-th layer features from the diffusion model, e is the empty text, and M is the total number of diffusion feature layers.

Aggregating Instance Semantic Features. We employ multiple sets of learnable queries, each corresponding to a specific layer of diffusion features. Specifically, for the i-th layer of diffusion features, we map them to key ($\mathbf{K}^i$) and value ($\mathbf{V}^i$) vectors, while a corresponding set of randomly initialized learnable queries is mapped to query ($\mathbf{Q}^i$) vectors:

$$\mathbf{Q}^i = q_{init}^i \mathbf{W}_{\mathcal{Q}}^i,\ \mathbf{K}^i = f_{sd}^i \mathbf{W}_{\mathcal{K}}^i,\ \mathbf{V}^i = f_{sd}^i \mathbf{W}_{\mathcal{V}}^i,\ \text{with } q_{init}^i \in \mathbb{R}^{r \times c} \tag{4}$$

where $\mathbf{W}_{\mathcal{Q}}^i$, $\mathbf{W}_{\mathcal{K}}^i$ and $\mathbf{W}_{\mathcal{V}}^i$ are linear projection matrices, q_{init}^i denotes randomly initialized learnable queries, c is the dimension of q_{init}^i, and r is the sequence length of q_{init}^i. Next, the layer-wise object agent queries are formed as follows:

$$\hat{q}_{ins}^i = \mathrm{FFN}\left(\frac{\exp\left(s^i\right)}{\sum_{j=1}^{hw}\exp\left(s^i\right)} \times \mathbf{V}^i\right),\ s^i = \frac{\mathbf{Q}^i(\mathbf{K}^i)^T}{\sqrt{d^i}},\ \text{with } \hat{q}_{ins}^i \in \mathbb{R}^{r \times c} \tag{5}$$

where $\hat{q}_{ins}^i$ are the object agent queries of the i-th diffusion layer, d^i is a scaling factor, and FFN consists of a linear mapping followed by an activation layer. h and w denote the height and width of the similarity matrix s^i, respectively. After the final layer, M, we compute both the maximum and average components across all layer-wise object agent queries to obtain the global object agent queries:

$$q_{ins} = \left(\max_{i=1,2,\ldots,M} \hat{q}_{ins}^i + \frac{1}{M}\sum_{i=1}^{M} \hat{q}_{ins}^i\right) \times W_a + b_a,\ \text{with } q_{ins} \in \mathbb{R}^{r \times c} \tag{6}$$

where W_a and b_a signify the weights and biases, respectively. Average queries encode global contextual information to improve robustness to noise, while max queries emphasize the most prominent and distinctive signals, highlighting key semantic features. As a result, the object agent queries establish implicit associations with scene instances and learn the semantic dependencies among objects embedded in diffusion features. Acting as an interface, these queries guide the 3D feature representations to move beyond isolated individuals, enabling the model to grasp consistent semantic patterns across diverse scenario domains.

3.4 Dual-queries Consistency Scheme

As diffusion models are inherently designed for generative tasks, their feature spaces encode not only high-level instance semantics but also intricate local visual details. These details, though irrelevant to 3D semantic segmentation, may be inadvertently propagated into object agent queries during interaction process and disrupt the learning of 3D semantic representations. To this end, we propose a dual-query consistency scheme to decouple potential local visual details from the object agent queries.

Dual-queries Design. To achieve this, we first construct an auxiliary set of object agent queries q'_{ins} guided by diffusion features at higher timesteps t', following the procedure described in Section 3.3. Different timesteps in the diffusion process correspond to successive stages of denoising. Higher timesteps introduce stronger noise that blurs local visual details, while the overall semantic structure of the scene remains largely preserved. As a result, the primary difference between the diffusion features at timestep t' and t lies in the superficial visual details.

Principal Component Extraction and Consistency Constraint. Recently, SoMA [65] reveals that the principal components derived from singular value decomposition (SVD) of weight matrices in vision foundation models (VFMs) capture generalized world knowledge, which underpins their strong generalization capability. Inspired by this, the object agent queries, as learnable parameters, are expected to exhibit similar properties with principal components that emphasize domain-invariant

instance semantic knowledge rather than intricate visual details. To this end, we first perform singular value decomposition (SVD) on the two sets of object agent queries:

$$q_{ins} = U\Sigma\mathcal{V}^T, \quad q'_{ins} = U'\Sigma'\mathcal{V}^{T,\prime}, \quad \text{with } U \in \mathbb{R}^{r\times r},\ \Sigma \in \mathbb{R}^{r\times c},\ \mathcal{V} \in \mathbb{R}^{c\times c} \tag{7}$$

where U (or U') and $\mathcal{V}$ (or $\mathcal{V}'$) are the left and right singular vectors, Σ (or Σ') is a diagonal matrix whose entries are singular values arranged in descending order. We select the top-k singular values along with their associated components in U (or U') and $\mathcal{V}$ (or $\mathcal{V}'$) to generate the quantized object agent queries q_{ins_pc} and q'_{ins_pc}:

$$q_{ins\_pc} = U_{[:,:k]}\Sigma_{[:k]}\mathcal{V}^T_{[:k,:]},\ q'_{ins\_pc} = U'_{[:,:k]}\Sigma'_{[:k]}\mathcal{V}^{T,\prime}_{[:k,:]},\ \text{with } q_{ins\_pc} \in \mathbb{R}^{r\times c},\ q'_{ins\_pc} \in \mathbb{R}^{r\times c} \tag{8}$$

Subsequently, we impose a consistency constraint between the two sets of quantized agent queries:

$$\mathcal{L}_{\text{consist}} = \sum_{\hat{r}=1}^{r} q'_{ins\_pc}(\hat{r}) \log \frac{q'_{ins\_pc}(\hat{r})}{\mathrm{MLP}(q_{ins\_pc})(\hat{r})} \tag{9}$$

where the MLP includes a linear transformation followed by layer normalization. By enforcing a consistency constraint, the principal components of the object agent queries are effectively guided to focus on instance semantic distribution rather than local visual details. This is because encoding excessive visual details within the principal components of the object agent queries would result in substantial discrepancies between the two sets of queries, q_{ins_pc} and q'_{ins_pc}. Thus, the consistency loss reduces these discrepancies, ensuring that only semantic dependencies among objects are retained in the constructed object agent queries, while irrelevant visual details are effectively suppressed.

3.5 Cross-modal Semantic Enhancement

Our ultimate goal is to improve the cross-domain generalization of 3D segmentation models. To this end, we leverage object agent queries as an interface for cross-modal semantic alignment, which subsequently enables the refinement of 3D feature representations to better capture semantic dependencies among objects across domains.

Cross-modal Semantic Alignment. Before that, we first generate the 3D object agent queries corresponding to the point cloud features. Similarly, we introduce multiple sets of learnable queries, each corresponding to a specific layer of the 3D backbone. These queries are projected into query vectors, while the associated 3D features are transformed into key and value vectors. A cross-attention mechanism is then employed to facilitate interaction between the learnable queries and 3D features, yielding layer-specific 3D agent queries. To generate the final 3D object agent query, we aggregate the outputs from all layers using both max and average pooling, followed by an MLP projection. For clarity, the previously introduced object agent queries derived from diffusion features are referred to as 2D object agent queries in the remainder of this paper. Next, we perform a dot product operation on the 2D object agent queries q_{ins_pc} with the 3D object agent queries q_{3D} to obtain a similarity map:

$$S = \mathrm{Softmax}\left(q_{3D} \times \mathrm{Linear}(q_{ins\_pc})^T\right), \quad \text{with } S \in \mathbb{R}^{r\times r} \tag{10}$$

where Linear is a two-layer MLP with layer normalization. The dot product operation produces a similarity matrix S, explicitly linking instance semantic information between the 2D and 3D object agent queries. Subsequently, we leverage S to recompose the semantic information from the 2D object agent queries, resulting in a new set of enhanced 3D object agent queries $\hat{q}_{3D}$ infused with the object-wise semantic dependencies conveyed by the 2D queries:

$$\hat{q}_{3D} = \mathrm{MLP}(S \times q_{ins\_pc} + q_{3D}), \quad \text{with } \hat{q}_{3D} \in \mathbb{R}^{r\times c} \tag{11}$$

Feature Refinement and Mask Prediction. The enhanced 3D object agent queries $\hat{q}_{3D}$ act as supervisory signals to guide the original 3D agent queries in actively modeling inter-instance semantic dependencies during their construction process. Inspired by UniDSeg [55] and xMUDA [19], we refine the segmentation features in the 3D decoder using $\hat{q}_{3D}$, while encouraging q_{3D} to mimic this refinement process:

$$\hat{f}^s_{3D\_aug} = \mathrm{MLP}(\mathrm{Softmax}(f^s_{3D} \times (\hat{q}_{3D})^T) \times \hat{q}_{3D}),\ \hat{f}^s_{3D} = \mathrm{MLP}(\mathrm{Softmax}(f^s_{3D} \times (q_{3D})^T) \times q_{3D})$$

$$\mathcal{L}^s_{\text{sup}} = \sum_{j=1}^{K} L_\delta(\hat{f}^s_{3D}, \hat{f}^s_{3D\_aug}), \quad \text{with } L_\delta(x,y) = \begin{cases} 0.5(x-y)^2 & \text{if } |x-y| < 1 \\ |x-y| - 0.5 & \text{otherwise} \end{cases} \tag{12}$$

Table 1: Performance comparison of domain adaptive and domain generalized 3D semantic segmentation methods in four typical settings. Top three results are highlighted as best, second and third, respectively. xM denotes the result which is obtained by taking the mean of the predicted 2D and 3D probabilities after softmax.

S:Source /T:Target		vKITTI/sKITTI		nuScenes:USA/Sing		nuScenes:Day/Night		A2D2/sKITT	
Task	Method	3D	xM	3D	xM	3D	xM	3D	xM
DA	logCORAL [34]	36.8	47.0	63.2	69.4	68.7	63.7	41.0	42.2
	MinEnt [50]	43.3	47.1	61.5	66.0	68.8	63.6	39.6	42.6
	BDL [29]	44.3	35.6	64.8	70.4	69.6	63.0	41.7	45.2
	xMUDA [19]	46.7	48.2	63.2	69.4	69.2	67.4	46.0	44.0
	AUDA [31]	37.8	41.3	64.0	69.2	69.8	64.8	43.6	46.8
	DsCML [37]	38.4	45.5	56.2	66.1	49.3	53.2	45.1	44.5
	Dual-Cross [27]	35.1	44.2	58.1	66.5	69.7	68.0	40.0	48.6
	SSE [67]	40.0	49.6	63.9	69.2	69.0	68.9	46.8	48.4
	BFtD [54]	45.5	51.5	62.2	69.4	70.4	68.3	44.4	48.7
	MM2D3D [5]	50.3	56.5	66.8	72.4	70.2	72.1	46.1	46.2
	VFMSeg [61]	52.0	61.0	65.6	72.3	70.5	66.5	52.3	50.0
	UniDSeg [55]	50.9	62.0	67.6	72.9	71.2	71.2	55.4	57.5
	XDiff3D	**53.1**	**63.3**	**69.5**	**74.1**	**73.5**	**72.3**	**57.6**	**58.8**
DG	xMUDA [19]	37.4	39.0	62.3	68.6	68.9	59.6	36.7	41.6
	MM2D3D [5]	40.2	44.2	62.3	70.9	63.2	68.3	35.9	43.6
	UniDSeg [55]	44.7	60.0	64.5	72.3	70.5	70.0	46.3	54.4
	XDiff3D	**46.9**	**61.3**	**66.7**	**73.5**	**72.4**	**71.6**	**49.1**	**56.2**

where f_{3D}^{s} denotes s-th stage features of the 3D decoder and $1 \leq s \leq S$. After the final stage, S, the refined segmentation features $\hat{f}_{3D}^{S}$ are fed into a 3D convolution to generate the final segmentation predictions:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{ce}} \left(\text{Conv}(\hat{f}_{3D}^{S}), Y \right) \tag{13}$$

where Y denotes the point cloud labels and $\mathcal{L}_{\text{ce}}$ denotes the cross-entropy loss.

Full Objective. Ultimately, the overall objective of the training process is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{seg}} + \gamma \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{consist}} \tag{14}$$

where γ and λ are hyperparameters. Notably, all operations involving 2D object agent queries and diffusion models are discarded during inference to ensure our framework remains concise and general.

4 Experiments

4.1 Datasets and Metrics

Following prior works [19, 58, 55, 68], we evaluate our method on six publicly available autonomous driving datasets, comprising three real-world datasets (nuScenes [13], SemanticKITTI [3], and A2D2 [15]), two synthetic datasets (VirtualKITTI [14] and SynLiDAR [57]), and one adverse-weather dataset (SemanticSTF [58]). The real-world datasets provide synchronized and calibrated LiDAR and RGB sensor data, enabling direct 2D-to-3D projection, whereas VirtualKITTI includes depth maps from which we simulate LiDAR scans by uniformly sampling points. For more details about the datasets, please refer to the supplementary material. Following standard practice [58, 55, 68], we evaluate segmentation performance using the mean Intersection over Union (mIoU) averaged over all classes for each dataset.

4.2 Implementation Details

Following prior works [55, 19], we adopt SparseConvNet [16] with a U-Net-style architecture as our 3D segmentation model, implemented using the Sparse Convolution Library [11]. The voxel resolution is set to 5cm, ensuring that each voxel encapsulates a single 3D point and provides sufficient spatial granularity for semantic segmentation. For the diffusion model, we leverage Stable Diffusion v2-1 [41], pretrained on the LAION-5B dataset [45], which is kept frozen during the entire training process. The model is optimized using AdamW [33], with a learning rate of 1e-5 for the 3D backbone

Table 2: Comparison of previous domain generalization methods on SemanticKITTI→SemanticSTF and SynLiDAR→SemanticSTF benchmarks.

Method	Dense-fog	Light-fog	Rain	Snow	Dense-fog	Light-fog	Rain	Snow
	SemanticKITTI→SemanticSTF				SynLiDAR→SemanticSTF			
Dropout [47]	29.3	25.6	29.4	24.8	15.3	16.6	20.4	14.0
Perturbation [58]	26.3	27.8	30.0	24.5	16.3	16.7	19.3	13.4
PolarMix [56]	29.7	25.0	28.6	25.6	16.1	15.5	19.2	15.6
MMD [26]	30.4	28.1	32.8	25.2	17.3	16.3	20.0	12.7
PCL [63]	28.9	27.6	30.1	24.6	17.8	16.7	19.3	14.1
PointDR [58]	31.3	29.7	31.9	26.2	19.5	19.9	21.1	16.9
UniMix [68]	34.8	30.2	34.9	30.9	24.3	22.9	26.1	20.9
XDiff3D	**37.5**	**34.1**	**38.1**	**33.7**	**26.2**	**24.6**	**27.9**	**22.5**

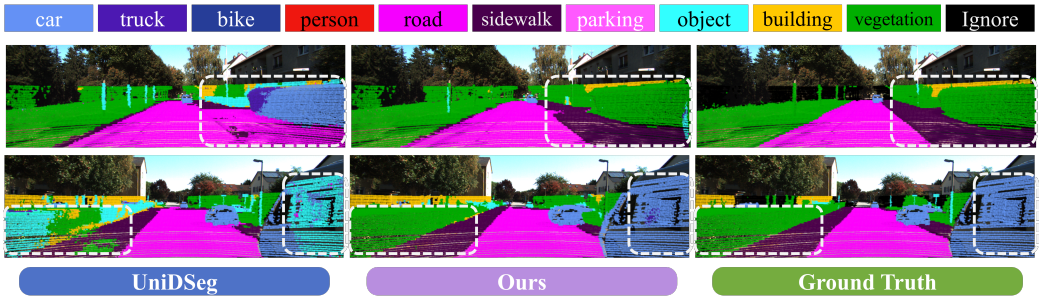


Figure 3: Qualitative results of DG3SS. From left to right: the visual results predicted by UniDSeg, Ours, and Ground Truth. We deploy the white dash boxes to highlight different prediction parts.

and 1e-4 for the decoder and learnable queries. Training is conducted for 50000 iterations with a batch size of 8. All experiments are performed on 4 NVIDIA RTX 4090 GPUs.

4.3 Comparison with State-of-the-art Methods

We comprehensively compare our method with existing domain adaptive 3D semantic segmentation (DA3SS) and domain generalized 3D semantic segmentation (DG3SS) methods. We conduct experiments in three standard evaluation settings: synthetic-to-real (VirtualKITTI→SemanticKITTI), real-to-real (nuScenes:USA→Sing, nuScenes:Day→Night, and A2D2→SemanticKITTI) and normal-to-adverse (SemanticKITTI→SemanticSTF and SynLiDAR→SemanticSTF). For the 2D branch architecture, we adopt the same structure as UniDSeg [55]. For DA3SS experiments, we also follow the exact configurations and training protocols of UniDSeg to ensure consistency and fair comparison.

Synthetic-to-real generalization. In Table 1, we compare our method with existing DG3SS and DA3SS approaches under the VirtualKITTI→SemanticKITTI setting. Our method significantly outperforms the previous state-of-the-art method UniDSeg in both 3D-only and 2D–3D fusion (i.e. xM) settings. Notably, in the 3D-only configuration, our method surpasses UniDSeg by 2.2 points in mIoU, demonstrating its strong cross-domain generalization capability.

Real-to-real generalization. In this experimental setting, models are evaluated under the real-to-real benchmarks, including nuScenes:USA→Sing, nuScenes:Day→Night, and A2D2→SemanticKITTI. As illustrated from the fifth column onward in Table 1, our method consistently achieves superior performance across all datasets in both DG and DA scenarios, surpassing previous state-of-the-art methods by clear margins. These results highlight the robustness and strong generalization capability of our framework in effectively handling complex and realistic domain variations encountered in practical 3D semantic segmentation tasks. Figure 3 further provides qualitative comparisons on the A2D2→SemanticKITTI benchmark, where our approach yields more complete and coherent predictions for road, sidewalk, building, and vegetation regions, effectively resolving ambiguous boundaries and reducing unreasonable predictions observed in UniDSeg. More qualitative results and detailed analyses are included in the supplementary material.

Normal-to-adverse generalization. In this experimental setting, all models are evaluated on the SemanticSTF dataset, which includes a range of challenging weather conditions. As shown in Table

Table 3: Ablation study on primary components, where 3D_AQ denotes 3D object agent queries.

	3D_AQ	$\mathcal{L}_{\text{sup}}$	$\mathcal{L}_{\text{consist}}$	nuScenes:USA/Sing		nuScenes:Day/Night		A2D2/sKITT	
				3D	xM	3D	xM	3D	xM
1				64.5	72.3	70.5	70.0	46.3	54.4
2	✓			64.9	72.5	70.7	70.1	46.8	54.7
3	✓	✓		66.0	73.1	71.6	71.2	48.0	55.6
4	✓	✓	✓	66.7	73.5	72.4	71.6	49.1	56.2

Table 4: Ablation study on different timestep selections under the DG3SS setting.

t	t'	nuScenes:USA/Sing		A2D2/sKITT	
		3D	xM	3D	xM
0	50	66.1	72.9	48.4	55.6
50	75	66.5	73.4	48.6	56.0
50	150	**66.7**	**73.5**	**49.1**	**56.2**
150	300	66.3	73.1	48.3	55.6
300	400	65.7	72.7	47.8	55.3

Table 5: Comparison of diffusion models under DG3SS on VirtualKITTI→SemanticKITTI.

Model	UniDSeg	+SD 1.4	+SD 1.5	+SD 2.1
mIoU	44.7	**46.1**	**46.6**	**46.9**

Table 6: Ablation on principal components under DG3SS on VirtualKITTI→SemanticKITTI.

k	20	50	70	90
mIoU	45.7	46.1	**46.9**	46.7

2, our method consistently outperforms existing approaches across all adverse scenarios. Notably, under the SemanticKITTI→SemanticSTF setting, our approach achieves a significant improvement over the previous state-of-the-art, exceeding it by more than 2 mIoU points on average. These results demonstrate the strong robustness and generalization ability of our method in the face of severe domain shifts.

Comparison of the class-wise IoU. We present a class-wise IoU analysis in Figure 4, using UniDSeg as the baseline model. The results reveal consistent improvements across a wide range of categories. The heatmap further illustrates the robustness of our method in capturing meaningful semantic structures and maintaining performance across domain shifts.

4.4 Ablation Study

This section presents comprehensive experimental results to verify the effectiveness of the proposed method. For more ablation studies and detailed analyses, please refer to the supplementary material.

Effect of components. We conduct ablation studies under DG3SS settings to validate the effectiveness of key components in our proposed method, specifically examining the contributions of the 3D object agent queries, $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{consist}}$. Using UniDSeg as our baseline model, the results presented in Table 3 reveal that: (1) Each component individually enhances performance, confirming their respective effectiveness. (2) Integrating all three components yields the highest mIoU scores across all benchmarks, underscoring their complementary roles in improving cross-domain segmentation generalization.

Study of the object agent query dimension c. The object agent queries serve as the core component of our framework. To assess the impact of their feature dimensionality, we experiment with values ranging from 64 to 1024. As shown in Figure 5, setting $c = 256$ yields competitive performance, achieving mIoU scores of 66.7% and 72.4% on the nuScenes:USA/Sing and nuScenes:Day/Night benchmarks, respectively.

The choice of t and t'. Different timesteps in the diffusion process correspond to different denoising stages, with larger timesteps introducing stronger noise. Prior studies [60, 2] indicate that effective timesteps typically fall within the range of 0 to 300, and that adjacent timesteps often yield highly similar features. To ensure sufficient diversity while maintaining semantic integrity, we select representative intervals from this range. As shown in Table 4, we empirically choose $t = 50$ and $t' = 150$ to effectively capture high-level semantic information while minimizing interference from low-level visual details.

Comparing different stable diffusion. As illustrated in Table 5, we adopt UniDSeg as the baseline and compare the performance of our method using three representative stable diffusion models. The results demonstrate that our method is robust to the choice of diffusion model, consistently outperforming the baseline and achieving significant performance gains.

Table 7: Effect of using different 3D backbones on the DA3SS methods.

3D Backbone	DA3SS	USA/Sing	
		3D	xM
SparseConvNet	UniDSeg	67.6	72.9
	XDiff3D	69.5	74.1
MinkowskiNet	UniDSeg	68.6	73.1
	XDiff3D	70.2	74.3

	Veget.	Build.	Road	Object	Truck	Car
Base	60.66	53.37	82.24	4.64	24.53	80.13
Ours	61.18	58.04	83.26	5.09	30.02	81.64

Figure 4: Comparison of the class-wise IoU on VirtualKITTI → SemanticKITTI under the DA3SS setting, using UniDSeg with and without our method.

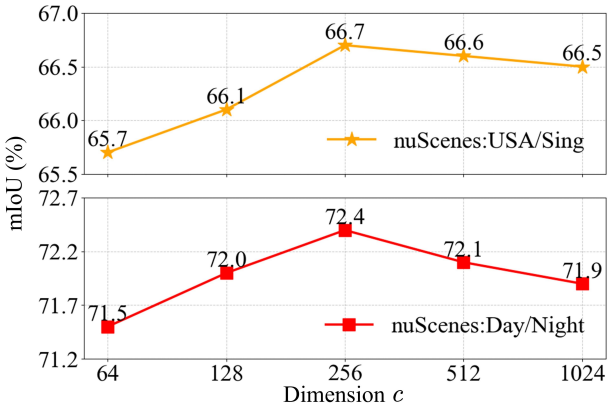


Figure 5: Ablation study on agent queries dimension c.

The number of principal components k. The results in Table 6 indicate that the selection of k—the number of retained principal components—significantly influences segmentation performance, with $k = 70$ achieving the highest mIoU. Increasing k beyond this value introduces fine-grained visual details, which interferes with the learning of domain-invariant semantics and leads to performance degradation.

Ablation on different backbones. As shown in Table 7, we evaluate the performance of various backbones integrated into our framework under consistent parameter settings. The results indicate that stronger backbones yield improved performance, and our method consistently outperforms the baseline across all configurations.

Table 8: Performance comparison under different noise levels.

σ (m)	0.05	0.1	0.2	0.3
UniDSeg (baseline)	40.3	37.1	33.8	30.6
Ours	**44.7**	**43.2**	**39.6**	**34.7**

Robustness to noisy point clouds. To evaluate robustness against measurement noise, we conducted an ablation study on A2D2 → SemanticKITTI benchmarks by adding zero-mean Gaussian noise with varying standard deviations σ. As shown in Table 8, our method consistently outperforms the UniDSeg baseline across all noise levels, maintaining higher mIoU even under severe perturbations. This demonstrates that our model possesses strong robustness and effectively preserves semantic consistency under noisy conditions.

5 Conclusion

In this work, we propose XDiff3D, a novel cross-modal framework guided by diffusion models to enhance the generalization of 3D semantic segmentation. XDiff3D constructs object agent queries to capture semantic dependencies among objects from diffusion features and infuses them into the 3D representation space. To mitigate interference from visual details, we further propose a dual-queries consistency scheme that encourages object-agent queries to focus on domain-invariant semantics. Extensive experiments on DG3SS and DA3SS benchmarks demonstrate that XDiff3D significantly outperforms previous SOTA methods, underscoring its effectiveness. This work addresses a significant gap in existing research and sets a new benchmark for cross-domain 3D semantic segmentation.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grant No.42504030 and No.52302506; the Fundamental Research Funds for the Central Universities (Science and Technology Program) under Grant No.D5000250047; and the Shaanxi Key Research and Development Program under Grant No.2025GH-YBXM-022.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022.

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=SlxSY2UZQT`.

[3] 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*, pages 9297–9307, 2019.

[4] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. Collaborating foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3108–3119, 2024.

[5] Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Exploiting the complementarity of 2d and 3d networks to address domain-shift in 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 98–109, 2023.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[7] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023.

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34: 17864–17875, 2021.

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[10] 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*, pages 3075–3084, 2019.

[11] Spconv Contributors. Spconv: Spatially sparse convolution library. *SpConv: Spatially sparse convolution library*, 2022.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[13] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.

[14] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[15] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

[16] 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*, pages 9224–9232, 2018.

[17] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational visual media*, 7:187–199, 2021.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[19] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1533–1544, 2022.

[20] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21741–21752, 2023.

[21] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023.

[22] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *European Conference on Computer Vision*, pages 91–109. Springer, 2024.

[23] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[24] Hyeonseong Kim, Yoonsu Kang, Changgyoon Oh, and Kuk-Jin Yoon. Single domain generalization for lidar semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17587–17598, 2023.

[25] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9338–9345. IEEE, 2023.

[26] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[27] Miaoyu Li, Yachao Zhang, Yuan Xie, Zuodong Gao, Cuihua Li, Zhizhong Zhang, and Yanyun Qu. Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3829–3837, 2022.

[28] Miaoyu Li, Yachao Zhang, Xu Ma, Yanyun Qu, and Yun Fu. Bev-dg: Cross-modal learning under bird's-eye view for domain generalization of 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11632–11642, 2023.

[29] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6936–6945, 2019.

[30] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1280–1289, 2022.

[31] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:211–221, 2021.

[32] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in neural information processing systems*, 32, 2019.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[34] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJWechg0Z`.

[35] Joshua Niemeijer, Manuel Schwonberg, Jan-Aike Termöhlen, Nico M Schmidt, and Tim Fingscheidt. Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2830–2840, 2024.

[36] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. In *European Conference on Computer Vision*, pages 37–54. Springer, 2024.

[37] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7108–7117, 2021.

[38] Duo Peng, Ping Hu, Qiuhong Ke, and Jun Liu. Diffusion-based image translation with label guidance for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 808–820, 2023.

[39] 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*, pages 652–660, 2017.

[40] 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, 2017.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[42] Kwonyoung Ryu, Soonmin Hwang, and Jaesik Park. Instant domain augmentation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9350–9360, 2023.

[43] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Elisa Ricci, and Fabio Poiesi. Cosmix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In *European Conference on Computer Vision*, pages 586–602. Springer, 2022.

[44] Jules Sanchez, Jean-Emmanuel Deschaud, and François Goulette. Domain generalization of 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18077–18087, 2023.

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[48] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.

[49] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024.

[50] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019.

[51] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28619–28630, 2024.

[52] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023.

[53] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023.

[54] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Cross-modal unsupervised domain adaptation for 3d semantic segmentation via bidirectional fusion-then-distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 490–498, 2023.

[55] Yao Wu, Mingwei Xing, Yachao Zhang, Xiaotong Luo, Yuan Xie, and Yanyun Qu. Unidseg: Unified cross-domain 3d semantic segmentation via visual foundation models prior. *Advances in Neural Information Processing Systems*, 37:101223–101249, 2024.

[56] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. *Advances in Neural Information Processing Systems*, 35:11035–11048, 2022.

[57] Aoran Xiao, Jiaxing Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2795–2803, 2022.

[58] Aoran Xiao, Jiaxing Huang, Weihao Xuan, Ruijie Ren, Kangcheng Liu, Dayan Guan, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9382–9392, 2023.

[59] Bowei Xing, Xianghua Ying, Ruibin Wang, Jinfa Yang, and Taiyan Chen. Cross-modal contrastive learning for domain adaptation in 3d semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2974–2982, 2023.

[60] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.

[61] Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Rui Zhang, Qingyuan Zhou, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *arXiv preprint arXiv:2403.10001*, 2024.

[62] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European conference on computer vision*, pages 677–695. Springer, 2022.

[63] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7097–7107, 2022.

[64] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022.

[65] Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[66] Hao Zhang, Feng Li, Huaizhe Xu, Shijia Huang, Shilong Liu, Lionel M Ni, and Lei Zhang. Mp-former: Mask-piloted transformer for image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18074–18083, 2023.

[67] Yachao Zhang, Miaoyu Li, Yuan Xie, Cuihua Li, Cong Wang, Zhizhong Zhang, and Yanyun Qu. Self-supervised exclusive learning for 3d segmentation with cross-modal unsupervised domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3338–3346, 2022.

[68] Haimei Zhao, Jing Zhang, Zhuo Chen, Shanshan Zhao, and Dacheng Tao. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14781–14791, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim our contribution, that is, a cross-modal learning framework to enhance the generalizability of cross-domain 3D semantic segmentation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation of the the work is detailed discussed in the Appendix A.4

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results with proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the necessary information needed to reproduce the main experimental results are fully stated in the Section 4 and Appendix A. This information ensures the understanding the results which support the main claims and conclusions of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All results are evaluated on the public 3D datasets. We will make the code publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 4.2, we illustrate the training details, including networks, learning rates, batch size, etc.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our experimental results do not contain statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section 4.2, we state that all experiments are conducted on four NVIDIA RTX 4090 GPUs, each equipped with 24GB of RAM.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper has no societal impact of the work performed. No AI ethics are involved, and no private data is involved.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.