2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Veila: Panoramic LiDAR Generation from a Monocular RGB Image

Youquan Liu[1], Lingdong Kong[2], Weidong Yang[1], Ao Liang[2], Jianxiong Gao[1], Yang Wu[3], Xiang Xu[4], Xin Li[5], Linfeng Li[6], Runnan Chen[7] and Ben Fei[8]

***Abstract*— Realistic and controllable panoramic LiDAR data generation is critical for scalable 3D perception in autonomous driving and robotics. Existing methods either perform unconditional generation with poor controllability or adopt text-guided synthesis, which lacks fine-grained spatial control. Leveraging a monocular RGB image as a spatial control signal offers a scalable and low-cost alternative, which remains an open problem. However, it faces three core challenges: (i) semantic and depth cues from RGB vary spatially, complicating reliable conditioning generation; (ii) modality gaps between RGB appearance and LiDAR geometry amplify alignment errors under noisy diffusion; and (iii) maintaining structural coherence between monocular RGB and panoramic LiDAR is challenging, particularly in image-LiDAR's non-overlap regions. To address these challenges, we propose *Veila*, a novel conditional diffusion framework that integrates: (i) a Confidence-Aware Conditioning Mechanism (CACM) that strengthens RGB conditioning by adaptively balancing semantic and depth cues according to their local reliability; (ii) Geometric Cross-Modal Alignment (GCMA) for robust RGB-LiDAR alignment under noisy diffusion; and (iii) Panoramic Feature Coherence (PFC) for enforcing global structural consistency across monocular RGB and panoramic LiDAR. Additionally, we introduce two metrics – Cross-Modal Semantic Consistency and Cross-Modal Depth Consistency – to evaluate alignment quality across modalities. Experiments on nuScenes, SemanticKITTI, and our proposed KITTI-Weather benchmark demonstrate that Veila achieves state-of-the-art generation fidelity and cross-modal consistency, while enabling generative data augmentation that improves downstream LiDAR semantic segmentation.**

I. Introduction

LiDAR point clouds are indispensable for 3D perception in autonomous driving and robotics, facilitating critical tasks such as 3D scene understanding [1], [2]. However, acquiring large-scale, high-quality LiDAR data in diverse environments is prohibitively expensive and time-consuming [3], [4]. This motivates the development of generative models to synthesize diverse, high-fidelity, and controllable LiDAR data as a cost-effective alternative.

Among image generative models [5], diffusion models [6] have emerged as a powerful paradigm for high-quality data synthesis due to their ability to model complex distributions and support fine-grained controllability [7], [8], [9], underscoring their potential as a paradigm for generating LiDAR data characterized by sparsity and irregular structures [10]. Recent explorations have extended diffusion models to LiDAR data generation, demonstrating promising initial results. For instance, LiDARGen [11] and RangeLDM [12] utilize diffusion models in range-view space but provide limited controllability, while Text2LiDAR [13] employs text-driven conditioning but lacks fine-grained spatial control. Despite recent progress, more flexible and fine-grained conditional LiDAR generations are required.

Image-conditioned LiDAR generation might be a promising alternative. Monocular cameras are widely deployed in cost-sensitive and modern autonomous platforms [14], enabling scalable and low-cost data collection. The captured RGB images provide the necessary semantic and depth-related cues, which are complementary to each other. Semantic information encodes object categories and scene layouts, facilitating scene-level understanding [15]. Meanwhile, monocular RGB images implicitly contain depth-related visual cues (e.g., perspective, occlusion, and scale) that serve as structural priors and depth ordering for 3D reconstruction [16]. This raises a critical question: **How can diffusion models generate a high-fidelity panoramic LiDAR scene from a monocular RGB image?**

Despite monocular RGB images providing a potential alternative to generate panoramic LiDAR, it has three key challenges: (1) How to combine both semantic and depth cues to construct a reliable conditioning signal? Semantic and depth cues extracted from the image exhibit complementary strengths but spatially varying reliability, i.e., semantic cues perform better in textured regions, while depth cues are more stable in geometrically structured or textureless areas. Relying on either cue alone is suboptimal; (2) Maintaining robust cross-modal alignment between RGB appearance and LiDAR geometry is non-trivial, due to their inherent modality gap and the progressive noise in intermediate diffusion stages, which often leads to correspondence collapse and structural distortion; (3) Ensuring structural consistency across a LiDAR panorama is particularly challenging in regions beyond the RGB field of view, where the absence of conditioning signals can lead to geometric drift or discontinuities.

Corresponding authors: W. Yang, R. Chen, and B. Fei.

[1]Y. Liu, J. Gao and W. Yang are with the College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China youquan164@gmail.com, jxgao22@m.fudan.edu.cn, wdyang@fudan.edu.cn

[2]L. Kong and A. Liang are with the School of Computing, National University of Singapore, Singapore lingdong.kong@u.nus.edu, a_liang@u.nus.edu

[3]Y. Wu is with Nanjing University of Science and Technology, Nanjing, China wuy98419@163.com

[4]X. Xu is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China xu_xiang@nuaa.edu.cn

[5]X. Li is with Shanghai AI Laboratory, Shanghai, China lixin@pjlab.org.cn

[6]L. Li is with ByteDance, Singapore linfeng.up@gmail.com

[7]R. Chen is with the University of Sydney, Sydney, Australia crunnan@gmail.com

[8]B. Fei is with The Chinese University of Hong Kong, Hong Kong, China benfei@cuhk.edu.hk

Fig. 1. Motivation for ***Veila***. The top row shows monocular RGB images used as conditions under various weather settings (sunny, night, fog, snow). The bottom row presents panoramic LiDAR scans generated by our framework, enabling controllable high-fidelity cross-weather synthesis.

To address these gaps, we propose ***Veila***, a novel conditional diffusion framework for generating high-fidelity panoramic LiDAR scenes from a monocular RGB image. It introduces three components: (1) a Confidence-Aware Conditioning Mechanism (CACM) adaptively integrates the strengths of semantic and depth information from RGB images to obtain a reliable conditioning signal; (2) a Geometric Cross-Modal Alignment (GCMA) module that leverages epipolar geometry [17] to maintain robust RGB-LiDAR alignment throughout noisy diffusion stages; (3) a Panoramic Feature Coherence (PFC) strategy that enforces structural consistency across the generated panoramic LiDAR. By simultaneously ensuring reliable conditioning, robust cross-modal alignment, and global panoramic coherence, Veila addresses the fundamental challenges of RGB-to-LiDAR generation, thereby producing panoramic LiDAR scenes that are faithful to the RGB input and structurally consistent across the full field of view. Moreover, since existing evaluation protocols lack metrics to assess the consistency between generated LiDAR and monocular RGB conditions, we propose two novel metrics called Cross-Modal Semantic Consistency (CM-SC) and Cross-Modal Depth Consistency (CM-DC), which quantitatively evaluate cross-modal alignment. Additionally, due to the current datasets containing limited RGB-LiDAR pairs under adverse weather conditions, the applicability to generate real-world scenarios remains unverified. Therefore, to explore the ability of generating adverse weather LiDAR scenes as illustrated in Fig. 1, we present the KITTI-Weather dataset, which contains scenes in clean, foggy, snowy, and nighttime settings. Extensive experiments on nuScenes, SemanticKITTI, and KITTI-Weather benchmark demonstrate that our framework achieves state-of-the-art performance in both fidelity and cross-modal consistency. Moreover, it significantly improves downstream LiDAR segmentation.

The main contributions of this paper are summarized:

- We propose ***Veila***, the first diffusion framework for generating panoramic LiDAR under the guidance of monocular RGB images, addressing key challenges including reliable control signal extraction, noisy cross-modal alignment, and global structural coherence.
- We design three novel modules: *i)* CACM for acquiring complementary semantic and depth conditioning signals; *ii)* GCMA module for maintaining robust RGB-LiDAR alignment under noisy diffusion; and *iii)* PFC strategy to enforce global spatial consistency.
- We introduce a modality consistency evaluation protocol comprising two novel metrics, CM-SC and CM-DC, along with the KITTI-Weather benchmark to assess LiDAR generation under adverse weather conditions.
- We demonstrate the effectiveness of our method through extensive experiments on SemanticKITTI, nuScenes, and KITTI-Weather, achieving the state-of-the-art fidelity and improving downstream LiDAR semantic segmentation by generative data augmentation.

II. Related Work

Denoising Diffusion Models. Diffusion models (DDMs) have emerged as a powerful generative paradigm for generative modeling, achieving state-of-the-art results across 2D vision and 3D domains. Early works [18], [19] performed denoising directly in pixel space for high-quality image synthesis. To improve efficiency, Latent Diffusion Models (LDMs) [20] operate in perceptually compressed latent spaces. Transformer-based DDMs (e.g., DiT [21]) scale diffusion to higher resolutions with enhanced modeling capacity. Recent efforts extend DDMs to controllable generation [22] and 3D-aware tasks for 3D point cloud synthesis [23], [24], demonstrating their versatility for complex modalities like LiDAR.

LiDAR Scene Generation. Generative models for LiDAR data remain relatively underexplored. LiDARGen [11] pioneered diffusion-based LiDAR generation by learning the score function in range-view space. Building on this, LiDM [25] improved LiDAR geometry using latent diffusion with structural preservation. R2DM [26] refined diffusion architectures and analyzed fidelity-critical components for unconditional scene generation, while UltraLiDAR [27] adopted a VQ-VAE [28] framework for LiDAR completion. RangeLDM [12] targeted real-time efficiency, and Text2LiDAR [13] introduced text-driven control for scene synthesis. However, prior works primarily address unconditional generation or text conditioning, monocular RGB-conditioned LiDAR generation remains largely unexplored.

III. Methodology

In this section, we present ***Veila***, a conditional diffusion framework for panoramic LiDAR generation from a monocular RGB image as shown in Fig. 2. Our ***Veila*** comprises

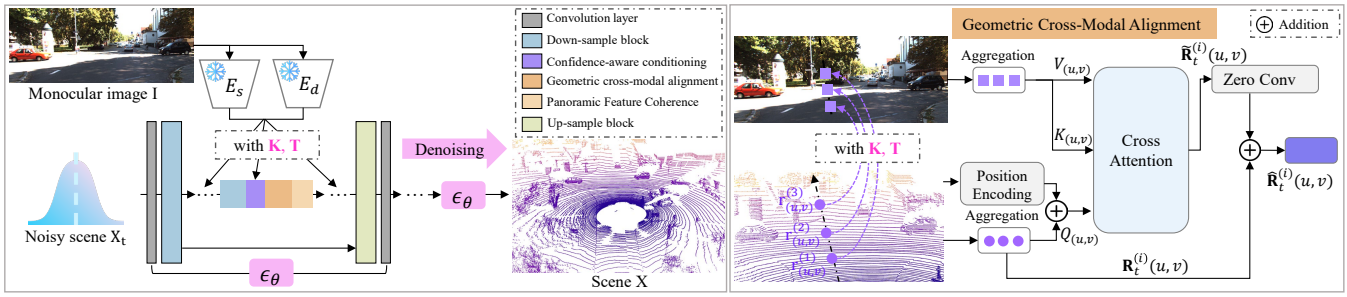

Fig. 2. **Method overview.** (*left*) Given a noisy scene X_t at diffusion step t, our framework progressively denoises it into a panoramic LiDAR scene X, conditioned on a monocular RGB image I. Three key designs are embedded within the U-Net backbone of the diffusion model: i) Confidence-Aware Conditioning Mechanism (CACM) adaptively integrates semantic and depth features from frozen encoders E_s and E_d; ii) Geometric Cross-Modal Alignment (GCMA) ensures robust alignment between RGB and LiDAR domains using the known camera matrices $\mathbf{K}$ and $\mathbf{T}$; and iii) Panoramic Feature Coherence (PFC) enforces global structural consistency across the panorama. (*right*) Illustration of GCMA: we show the LiDAR ray corresponding to a range image coordinate (u, v) as an example, highlighting how GCMA leverages epipolar geometry to facilitate cross-modal alignment under noisy diffusion stages.

three core components: a Confidence-Aware Conditioning Mechanism, a Geometric Cross-modal Alignment module, and a Panoramic Feature Coherence strategy, addressing key challenges in RGB-conditioned LiDAR generation. The following subsections describe each component in detail.

A. Preliminaries

Range Image Representation. A LiDAR point cloud is defined as $\mathcal{P} = \{(p^m, e^m) \mid m = 1, ..., N\}$, where each point $p^m \in \mathbb{R}^3$ represents 3D coordinates, and $e^m \in \mathbb{R}^L$ denotes auxiliary L attributes such as intensity. Following prior works [26], [13], we transform $\mathcal{P}$ into a structured range image $\mathrm{X} \in \mathbb{R}^{h \times w \times 2}$ via spherical projection [29]:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan(p_y^m, p_x^m)\pi^{-1}\right] w \\ \left[1 - (\arcsin(p_z^m d^{-1}) + f_{\text{down}}) f^{-1}\right] h \end{pmatrix}, \tag{1}$$

where (u, v) are range image coordinates, (h, w) are the height and width of the range image, $f = |f_{\text{up}}| + |f_{\text{down}}|$ is the vertical field-of-view of the sensor, and $d = \|p^m\|_2$ is the Euclidean distance. Each pixel encodes depth and intensity values, allowing standard 2D convolutional architectures to process 3D point cloud data effectively.

LiDAR-Camera Projection. To establish the alignment between LiDAR point clouds and camera image pixels, we project 3D LiDAR coordinates into 2D image coordinates. Given a 3D point $p^m = (p_x^m, p_y^m, p_z^m)$ in the LiDAR coordinate system, the corresponding RGB image coordinate (u', v') are obtained by:

$$[u', v', 1]^\top = \frac{1}{p_z^m} \cdot \mathbf{KT} \cdot [p_x^m, p_y^m, p_z^m, 1]^\top, \tag{2}$$

where $\mathbf{T} \in \mathbb{R}^{4\times4}$ denotes the camera extrinsic matrix and $\mathbf{K} \in \mathbb{R}^{3\times4}$ is the camera intrinsic matrix, and $[\cdot]^\top$ denotes the matrix transpose operation.

B. Problem Formulation

We regard RGB-conditioned panoramic LiDAR generation as a conditional diffusion process. Given a monocular RGB image $\mathrm{I} \in \mathbb{R}^{H\times W\times3}$ as condition, our goal is to generate a corresponding panoramic LiDAR point cloud represented as range image $\mathrm{X} \in \mathbb{R}^{h\times w\times2}$, where each pixel encodes depth and intensity. Following the denoising diffusion probabilistic model (DDPM) [19] framework, we learn to predict the Gaussian noise $\boldsymbol{\epsilon}$ added to clean data X_0 through a noise prediction network $\boldsymbol{\epsilon}_\theta$. The training objective minimizes:

$$\mathcal{L} = \mathbb{E}_{\mathrm{X}_0, \boldsymbol{\epsilon}\sim\mathcal{N}(0,1), t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathrm{X}_t, \mathrm{I}, t)\|^2\right], \tag{3}$$

where X_t is the noisy sample at timestep t. This formulation presents key challenges in constructing reliable conditioning signals, maintaining cross-modal alignment, and enforcing global structural coherence, which are addressed by the components introduced in the following sections.

C. Confidence-Aware Conditioning

RGB images provide both semantic and depth cues, which offer complementary guidance but vary in spatial reliability. Relying on either alone leads to suboptimal conditioning. To this end, we introduce CACM, which adaptively fuses semantic and depth features based on their local reliability.

Semantic and Depth Feature Extraction. Semantic information encodes object categories, scene layout, and contextual relationships, while depth cues reflect depth ordering and structural priors. To extract these complementary signals, we employ two parallel frozen encoders: a semantic encoder E_s and a depth encoder E_d, both pretrained on large-scale vision datasets. Freezing these encoders preserves pretrained knowledge and mitigates overfitting on limited LiDAR data. The semantic encoder E_s produces multi-scale feature maps $\{F_s^{(i)}\}_{i=1}^4$, with each $F_s^{(i)} \in \mathbb{R}^{H_i\times W_i\times D_i^s}$ capturing semantic context at scale i. Similarly, the depth encoder E_d outputs depth-aware features $\{F_d^{(i)}\}_{i=1}^4$, where $F_d^{(i)} \in \mathbb{R}^{H_i\times W_i\times D_i^d}$ encodes geometric structure.

Adaptive Conditioning from RGB Features. Although semantic and depth features provide complementary information, their spatial reliability varies: semantic cues are more informative in textured and semantically rich regions, while depth estimates are more stable in geometrically structured or textureless areas. This spatial complementarity indicates that neither cue alone is sufficient. Naive fusion strategies such as concatenation fail to capture this heterogeneity, leading to suboptimal conditioning. In diffusion-based generation, such inconsistencies may accumulate across denoising steps,

resulting in structural distortions or semantic artifacts. To address this, we propose a Confidence-Aware Conditioning Mechanism that adaptively integrates semantic and depth cues based on local reliability. Since semantic features preserve finer spatial detail, we first interpolate depth features to match the semantic resolution and project both to a shared latent space:

$$\tilde{F}_s^{(i)} = \text{Conv}(F_s^{(i)}), \quad \tilde{F}_d^{(i)} = \text{Conv}(\text{Interp}(F_d^{(i)})), \tag{4}$$

where $\text{Interp}(\cdot)$ denotes bilinear interpolation. Confidence scores are computed through separate estimators analyzing local feature statistics:

$$c_s^{(i)} = \text{Sigmoid}(\text{Conv}(\tilde{F}_s^{(i)})), \quad c_d^{(i)} = \text{Sigmoid}(\text{Conv}(\tilde{F}_d^{(i)})). \tag{5}$$

The conditioning image features combine both cues with normalized confidence weighting:

$$F^{(i)} = \frac{c_s^{(i)} \cdot \tilde{F}_s^{(i)} + c_d^{(i)} \cdot \tilde{F}_d^{(i)}}{c_s^{(i)} + c_d^{(i)} + \delta}, \tag{6}$$

where δ ensures numerical stability. This mechanism allows the network to emphasize more reliable cues in each region, resulting in robust and spatially adaptive conditioning for LiDAR generation.

D. Geometric Cross-Modal Alignment

The key challenge in RGB-conditioned LiDAR generation lies in establishing accurate correspondences between RGB pixels and range image elements, especially under the noise corruption inherent in diffusion. Conventional projection-based methods depend on current 3D point locations, which become unreliable during intermediate denoising steps [30]. To address this, we propose a Geometric Cross-Modal Alignment module that leverages epipolar constraints derived from geometry-defined LiDAR ray directions (determined solely by the range image coordinates), enabling stable and noise-tolerant correspondences throughout the diffusion process. Specifically, each range image pixel (u, v) defines a deterministic ray direction from the LiDAR origin, derived from sensor geometry. Then, we compute the corresponding 3D coordinates $\mathbf{r}_{(u,v)}^{(k)}$ by applying $\text{Ray}(u, v)$, the inverse of Equation (1), which recovers the unit 3D point determined by the LiDAR inclination and azimuth angles. These unit points are then scaled by depths $\{d_k\}_{k=1}^{\mathcal{K}}$ to obtain the sampled 3D coordinates:

$$\mathbf{r}_{(u,v)}^{(k)} = d_k \cdot \text{Ray}(u, v). \tag{7}$$

These sampled 3D points $\mathbf{r}_{(u,v)}^{(k)}$ are projected to RGB image coordinates using Equation (2). Features are retrieved from the conditioning representations $F^{(i)}$ at the projected locations (u'_k, v'_k) and aggregated using depth-aware weights:

$$V_{(u,v)} = \frac{\sum_{k=1}^{\mathcal{K}} w_k \cdot F^{(i)}(u'_k, v'_k)}{\sum_{k=1}^{\mathcal{K}} w_k}, \tag{8}$$

where $w_k = \exp(-d_k/\tau) \cdot m_k$ combines exponential depth decay with validity masks m_k. To enrich the current range image features, we incorporate Fourier positional embeddings of the 3D coordinates $\mathbf{r}_{(u,v)}^{(k)}$. This encoding captures high-frequency spatial variations, enhancing geometric consistency in cross-modal alignment.

$$Q_{(u,v)} = \mathbf{R}_t^{(i)}(u, v) + \text{MLP}(\gamma(\mathbf{r}_{(u,v)}^{(k)})), \tag{9}$$

where $\gamma(\cdot)$ denotes the Fourier positional encoding function applied to 3D coordinates. $\mathbf{R}_t^{(i)}(u, v)$ represents the range image feature at timestep t and scale i. The cross-attention output is computed as:

$$\tilde{\mathbf{R}}_t^{(i)}(u, v) = \text{Softmax}\left(\frac{Q_{(u,v)} K_{(u,v)}^\top}{\sqrt{d_h}}\right) V_{(u,v)}, \tag{10}$$

where $K_{(u,v)}$ is set to $V_{(u,v)}$ and d_h is the attention head dimension. The output features $\hat{\mathbf{R}}_t^{(i)}(u, v)$ are integrated via zero-initialized convolutions with residual connections, as illustrated in Figure 2. This design preserves diffusion dynamics while progressively injecting RGB-guided alignment signals.

E. Panoramic Feature Coherence

Monocular RGB conditioning primarily affects front-view LiDAR regions where image observations are available. In contrast, rear-view regions lie outside the RGB field of view and therefore lack conditioning signals. During diffusion, these unobserved areas are effectively treated as unconditional generations, often resulting in structural discontinuities or semantic drift across the panoramic scene. This challenge is amplified by the limited receptive field of standard UNet-based diffusion architectures, which struggle to propagate global context across spatially distant regions [31]. Without explicit regularization, model may produce inconsistent object structures between conditioned and unconditioned views.

To mitigate this, we propose a Panoramic Feature Coherence strategy that introduces a global self-attention layer at the deepest stage of the UNet, where feature maps have the largest receptive field and capture high-level scene semantics. This component facilitates long-range dependencies, allowing semantic and geometric information from RGB-conditioned regions to propagate throughout the panorama:

$$\bar{\mathbf{R}}_t^{(d)} = \text{SelfAttention}(\hat{\mathbf{R}}_t^{(d)}) + \hat{\mathbf{R}}_t^{(d)}, \tag{11}$$

where d is the deepest UNet layer. This design promotes spatial coherence across the full panoramic field, helping maintain consistent structures even in unobserved regions. Meanwhile, local fidelity in RGB-visible areas is preserved.

IV. Experiments

A. Experimental Settings

Datasets. We conduct experiments on the nuScenes [4] and SemanticKITTI [3] datasets. To evaluate generalization under adverse weather, we introduce **KITTI-Weather**, a new benchmark constructed by augmenting KITTI [32] with adverse weather conditions. Specifically, clean RGB-LiDAR pairs from KITTI are used as base scenes, and night, fog, and snow variants are generated using **DriveGEN**-synthesized

TABLE I

COMPARISON OF **LIDAR SCENE GENERATION** METHODS ON THE *SemanticKITTI* DATASET. LOWER IS BETTER FOR ALL METRICS (↓).

Method	Venue	FRD↓	FPD↓	JSD↓	MMD↓
LiDARGen	ECCV'22	735.49	119.69	0.13	21.90
LiDM	CVPR'24	-	496.78	0.08	9.20
R2DM	ICRA'24	262.85	12.06	0.03	0.89
Text2LiDAR	ECCV'24	567.47	16.78	0.08	4.24
Veila	**Ours**	**220.40**	**8.18**	**0.02**	**0.72**

TABLE II

COMPARISONS OF **LIDAR SCENE GENERATION** METHODS ON THE *nuScenes* DATASET. METRICS WITH (↓) INDICATE LOWER IS BETTER.

Method	Venue	FRD↓	FPD↓	JSD↓	MMD↓
LiDARGen	ECCV'22	549.18	22.80	0.04	0.76
LiDM	CVPR'24	-	30.77	0.07	3.86
R2DM	ICRA'24	253.80	14.35	**0.03**	**0.48**
Text2LiDAR	ECCV'24	953.18	147.48	0.09	12.50
Veila	**Ours**	**232.37**	**12.11**	0.05	0.62

RGB images [33] and **Robo3D**-simulated LiDAR point clouds [34]. This design enables evaluation of LiDAR generation models in challenging weather scenarios, where such data is scarce. KITTI-Weather contains 3712 clean RGB-LiDAR pairs, along with 3712 pairs for each adverse-weather condition (night, fog, and snow), resulting in a balanced design across weather types.

Evaluation Metrics. Following prior work [26], [35], [36], we evaluate generation quality using four standard metrics: Fréchet Range Distance (FRD) and Fréchet Point Cloud Distance (FPD), which measure feature-level fidelity in the range image and point cloud domains, respectively; and Jensen-Shannon Divergence (JSD) and Maximum Mean Discrepancy (MMD), which evaluate distribution-level similarity between generated and real point clouds. MMD results are reported in 10^{-4} throughout this paper. To assess the cross-modal alignment between the generated LiDAR and the RGB input used for conditioning, we introduce two novel metrics: **Cross-Modal Semantic Consistency (CM-SC)** and **Cross-Modal Depth Consistency (CM-DC)**. CM-SC measures semantic alignment by assigning pseudo-labels to generated LiDAR points with a pretrained SPVCNN [37], projecting them onto the RGB image plane, and computing pixel-level accuracy against SegFormer [38] predictions after mapping both outputs into a unified label space. CM-DC evaluates depth consistency by projecting generated LiDAR points into the camera coordinate system and comparing their depths with monocular predictions from DepthAnything-V2 [16] using a root mean squared log error. These metrics evaluate both the structural realism of the generated LiDAR and its semantic and geometric consistency with the RGB input. For quantitative evaluation, we generated 5,000 samples on SemanticKITTI, 10,000 samples on nuScenes, and 1,000 samples on KITTI-Weather.

Implementation Details. All experiments were conducted on a single NVIDIA A40 GPU. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4.

TABLE III

COMPARISONS OF **LIDAR SCENE GENERATION** METHODS UNDER DIVERSE WEATHER CONDITIONS IN THE *KITTI-Weather* DATASET.

Type	Method	Venue	FPD↓	JSD↓	MMD↓
Clean/Night	LiDARGen	ECCV'22	484.48	0.14	18.89
	LiDM	CVPR'24	127.47	0.17	31.37
	Text2LiDAR	ECCV'24	23.61	0.06	37.63
	R2DM	ICRA'24	20.27	**0.05**	12.46
	Veila (clean)	**Ours**	**13.16**	0.07	**6.30**
	Veila (night)	**Ours**	**16.31**	0.07	**6.89**
Fog	LiDARGen	ECCV'22	557.63	0.17	25.31
	LiDM	CVPR'24	349.75	0.16	26.30
	Text2LiDAR	ECCV'24	40.54	0.13	19.83
	R2DM	ICRA'24	30.27	**0.10**	**11.86**
	Veila	**Ours**	**17.57**	**0.10**	14.83
Snow	LiDARGen	ECCV'22	484.48	0.16	27.66
	LiDM	CVPR'24	176.73	0.13	16.70
	Text2LiDAR	ECCV'24	23.69	0.05	14.32
	R2DM	ICRA'24	19.37	0.10	4.10
	Veila	**Ours**	**18.46**	**0.02**	**3.30**

TABLE IV

COMPARISON OF **LIDAR SCENE GENERATION** METHODS, WITH LIDAR SCENES PARTITIONED INTO FRONT-VIEW ($p_x^m > 0$) AND REAR-VIEW ($p_x^m \leq 0$) REGIONS ON *SemanticKITTI*.

Method	Venue	$p_x^m > 0$			$p_x^m \leq 0$		
		FPD↓	JSD↓	MMD↓	FPD↓	JSD↓	MMD↓
LiDARGen	ECCV'22	173.72	0.14	46.51	62.34	0.13	44.84
LiDM	CVPR'24	466.44	0.08	16.94	200.05	0.09	16.93
R2DM	ICRA'24	11.57	0.03	0.89	12.52	**0.04**	2.37
Text2LiDAR	ECCV'24	7.87	0.07	8.29	9.61	0.08	9.44
Veila	**Ours**	**4.81**	**0.02**	**0.74**	**7.57**	**0.04**	**2.19**

The models were trained for $400{,}000$ steps in total. During inference, 256 denoising steps were employed to balance efficiency and generation quality. For nuScenes [4], LiDAR range images were projected at a resolution of 32×1024, while SemanticKITTI and KITTI-Weather followed their native resolution of 64×1024. For CACM, two frozen encoders (E_s and E_d) were used to extract semantic and depth cues from monocular RGB images: SegFormer [38] pretrained on Cityscapes [39] for semantics, and DepthAnything-V2 [16] pretrained on Virtual KITTI 2 [40] for depth. Keeping these encoders fixed preserved pretrained knowledge and avoided overfitting on the limited training data. In GCMA, we sampled rays across eight denoiser scales in a symmetric configuration [48, 36, 24, 18, 18, 24, 36, 48], balancing efficiency and accuracy across feature resolutions.

To stabilize training, we normalize intensity and depth separately: intensity values are linearly scaled to $[-1, 1]$. Following R2DM [26], depth values are log-transformed and scaled to $[-1, 1]$, and angular coordinates are Fourier-encoded [26] to provide explicit geometric priors. For KITTI-Weather, each batch contains samples from only one weather type to ensure stable batch-normalization statistics.

B. Quantitative Results

LiDAR Scene Generation. We evaluate our method on SemanticKITTI, nuScenes, and KITTI-Weather against recent state-of-the-art approaches. Across all benchmarks (Table I-III), our method achieves strong improvements in FRD and

FPD over prior works, while maintaining competitive JSD and MMD. On SemanticKITTI (Table I), our approach reduces FRD by **16.1%** and FPD by **32.1%** relative to R2DM, demonstrating improved range-view fidelity and point-wise accuracy. On nuScenes (Table II), it achieves the **lowest FRD and FPD**, indicating robustness under sparse LiDAR settings. In adverse weather scenarios (Table III), our method maintains strong performance across clean, night, fog, and snow conditions. Notably, it achieves a **10–25% reduction** in FPD compared to R2DM and Text2LiDAR, demonstrating its robustness towards adverse weather and domain shifts. These consistent gains highlight the effectiveness of our conditioning and alignment designs for robust, high-fidelity, and structurally coherent LiDAR generation.

Region-Wise Generation Quality Assessment. To better assess the impact of RGB conditioning, we partition the generated panoramic LiDAR into front-view regions visible to the RGB camera ($p_x^m > 0$) and rear-view regions outside its field of view ($p_x^m \leq 0$). As shown in Table IV, our method achieves superior performance in both regions. While front-view areas benefit directly from RGB guidance, our framework also maintains high fidelity in rear-view regions where no image cues are available. This demonstrates the model's ability to propagate contextual information beyond the visible region, enabled by our PFC strategy.

Generative Data Augmentation. To evaluate the utility of our method for downstream LiDAR segmentation, we generate 10,000 synthetic frames per approach and assign pseudo-labels using a pretrained SPVCNN [37]. These frames are combined with 1%, 10%, and 20% subsets of the SemanticKITTI train set to train two segmentation backbones: MinkUNet [41] (voxel-based) and SPVCNN (voxel-point fusion). As shown in Table V, Veila achieves mIoU improvements of +1.6 and +1.3 over R2DM with MinkUNet and SPVCNN, respectively, in the challenging low-data (1%) setting. We attribute these gains to the superior fidelity and cross-modal consistency of our generated LiDAR, which not only produces visually realistic scenes but also provides more informative training signals for downstream perception.

C. Qualitative Results

To complement the quantitative results, we present qualitative examples demonstrating our framework's ability to generate high-fidelity LiDAR scenes, enable controllable scene editing, and maintain cross-modal consistency (Fig. 3–6).

Fig. 3 shows a qualitative comparison on the SemanticKITTI dataset. LiDARGen produces noisy and scattered point clouds with noticeable structural artifacts, while R2DM yields relatively sparse outputs that lack fine-grained detail. In contrast, our method generates realistic and coherent LiDAR scenes with clearly defined structures (e.g., cars), demonstrating superior fidelity and spatial completeness. Fig. 4 showcases results on KITTI-Weather. Compared to Text2LiDAR and R2DM, our framework produces LiDAR scenes that are more structurally consistent with the weather-degraded ground truth, better preserving fine details under foggy and snowy conditions. Fig. 5 demonstrates our method's capacity for controllable scene editing, including object removal, replacement, and weather modification. The edited LiDAR scenes maintain structural plausibility and semantic coherence, reflecting the model's strong geometric understanding across diverse conditions. Fig. 6 depicts the alignment between generated LiDAR and RGB conditions. Highlighted regions show precise object-level alignment, where the generated LiDAR point clouds accurately match object boundaries in the RGB conditions.

TABLE V

DOWNSTREAM APPLICATION OF ***Veila*** TO **LIDAR SEMANTIC SEGMENTATION** TASK ON THE *val* SET OF *SemanticKITTI*.

Base	Method	Venue	SemanticKITTI 1%	10%	20%
MinkUNet	*Sup.-only*	-	40.39	60.90	62.84
	LiDARGen	ECCV'22	36.11	54.73	60.39
	R2DM	ICRA'24	53.38	60.78	62.57
	Text2LiDAR	ECCV'24	40.23	55.00	58.35
	Veila	**Ours**	**55.01**	**61.25**	**63.00**
SPVCNN	*Sup.-only*	-	37.86	59.07	61.16
	LiDARGen	ECCV'22	36.44	55.04	59.71
	R2DM	ICRA'24	50.25	60.11	62.34
	Text2LiDAR	ECCV'24	40.55	53.87	58.34
	Veila	**Ours**	**51.53**	**60.40**	**62.71**

TABLE VI

ABLATION STUDY ON *SemanticKITTI* EVALUATING THE CONTRIBUTION OF EACH COMPONENT IN ***Veila***.

#	Configuration	FRD↓	FPD↓	JSD↓	MMD↓	CM-DC↓	CM-SC↑
a	E_s only	301.57	21.65	0.04	1.67	0.28	60.49
b	E_d only	281.25	27.55	0.03	1.01	0.27	58.15
c	Ours w/o CACM	342.26	17.94	0.072	4.68	0.25	58.34
d	Ours w/o GCMA	442.18	26.59	0.09	8.35	0.34	45.48
e	Ours w/o PFC	241.59	11.70	0.06	2.37	**0.24**	63.16
f	**Full Framework**	**220.40**	**8.18**	**0.02**	**0.72**	**0.24**	**63.92**

TABLE VII

COMPARISON OF SINGLE-VIEW AND MULTI-VIEW RGB CONDITIONING IN ***Veila*** ON THE *nuScenes*.

#	#Camera(s)	FRD↓	FPD↓	JSD↓	MMD↓	CM-DC↓	CM-SC↑
a	1	232.37	12.11	**0.05**	0.62	0.34	53.77
b	3	**215.20**	**10.40**	**0.05**	**0.56**	**0.29**	**58.69**

D. Ablation Study

Effects of Each Component. We evaluate the contribution of each module and feature configuration in ***Veila***. As shown in Table VI, using only semantic features from E_s or only depth features from E_d results in consistent degradation across all metrics. This highlights the complementary nature of the two cues: semantic features lack the geometric information required for structural fidelity, while depth features fail to capture semantic context. Removing CACM (w/o CACM) leads to increased FRD and FPD, illustrating the limitations of naive feature concatenation in balancing local semantic and geometric cues. The adaptive weighting in CACM allows the model to dynamically adjust the contributions of semantic and depth features based on their spatial reliability, resulting in more effective conditioning in heterogeneous regions. Excluding GCMA (w/o GCMA) causes a substantial

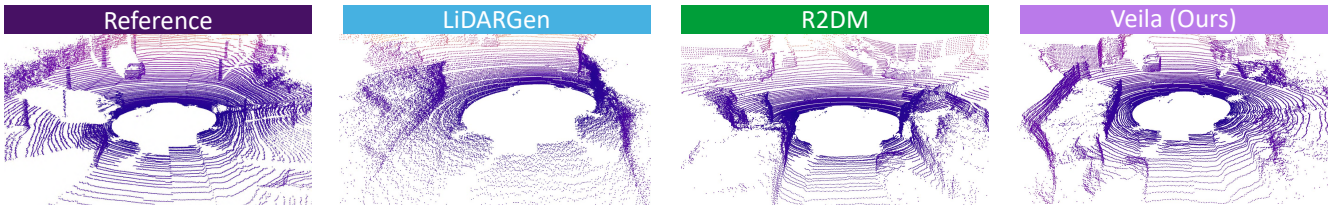

Fig. 3. Qualitative comparisons of ***Veila*** against state-of-the-art LiDAR scene generation approaches on the SemanticKITTI dataset. From left to right: Reference (ground truth), LiDARGen, R2DM, and our method.

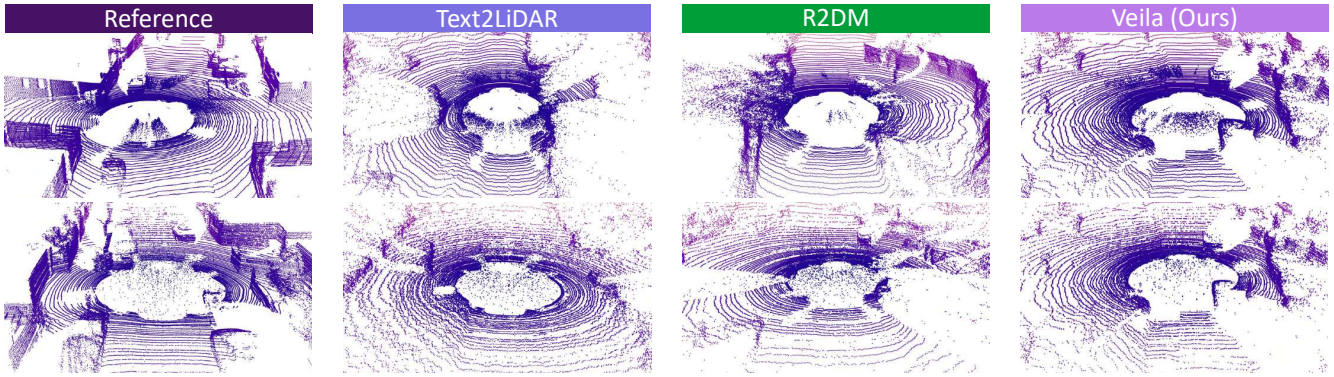

Fig. 4. Qualitative results on KITTI-Weather. Each row shows LiDAR scenes in different weather conditions: (top) foggy scenes and (bottom) snowy scenes. From left to right: Reference (ground truth), Text2LiDAR, R2DM, and our framework.

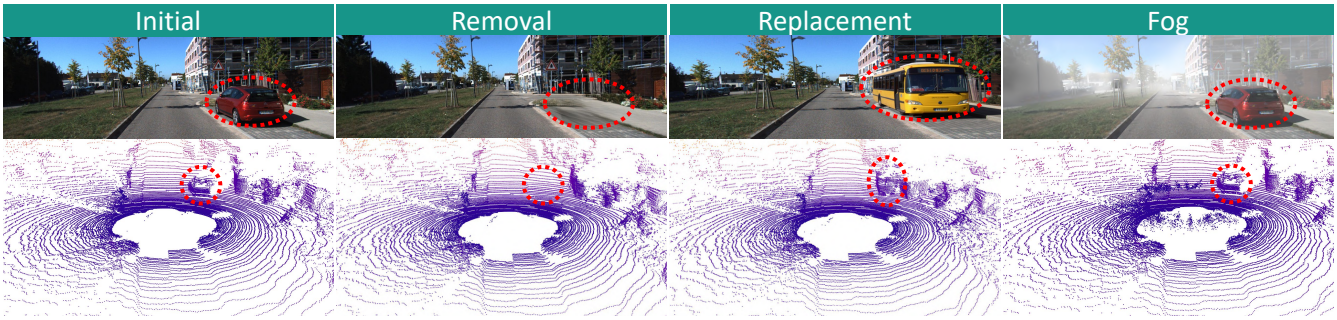

Fig. 5. Scene editing examples with ***Veila***. The highlighted regions mark edited objects and weather modification.

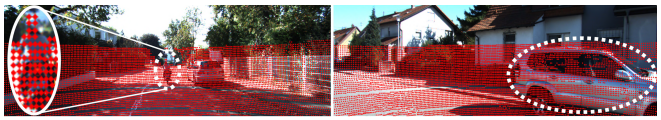

Fig. 6. LiDAR points generated from ***Veila*** are projected onto RGB images. Highlighted regions demonstrate precise object-level alignment.

drop in CM-SC and a rise in CM-DC, underscoring its role in maintaining RGB–LiDAR consistency. By enforcing robust geometric alignment between the modalities, GCMA helps mitigate cross-modal misalignment, even under noisy diffusion. Finally, removing PFC (w/o PFC) degrades global metrics such as FRD and FPD, indicating that the PFC strategy promotes structural coherence across the panoramic LiDAR, particularly in regions lacking RGB conditioning. Overall, the full framework achieves the best results across all evaluation metrics, confirming that CACM, GCMA, and PFC contribute complementary strengths to enable robust, high-fidelity panoramic LiDAR generation.

Impact of Additional Camera Inputs. To assess the benefit of multi-view RGB conditioning, we extend ***Veila*** to incorporate three camera views (front, front-left, and front-right) from *nuScenes*. Each view is independently encoded using E_s and E_d, and the resulting features are concatenated along the view dimension before being fed into our method. As shown in Table VII, incorporating side-view images improves both geometric fidelity and cross-modal consistency compared to using single-view conditioning. We attribute these gains to the expanded field of view and reduced occlusion, which provide richer contextual information.

V. Conclusion

We present ***Veila***, a novel diffusion framework designed for generating panoramic LiDAR from a monocular RGB image. By systematically addressing key challenges, our method enables realistic and controllable panoramic LiDAR generation. We further propose two cross-modal consistency evaluation metrics and introduce the KITTI-Weather benchmark to standardize assessment under adverse conditions. Extensive experiments on SemanticKITTI, nuScenes, and KITTI-Weather datasets show that our method achieves state-of-the-art fidelity and significantly enhances downstream LiDAR semantic segmentation.

REFERENCES

[1] R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, and W. Wang, "Clip2scene: Towards label-efficient 3d scene understanding by clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7020–7030, 2023.

[2] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu, "Segment any point cloud sequences by distilling vision foundation models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37193–37229, 2023.

[3] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. 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.

[4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. 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.

[5] L. Kong, W. Yang, J. Mei, Y. Liu, A. Liang, D. Zhu, D. Lu, W. Yin, X. Hu, M. Jia, *et al.*, "3d and 4d world modeling: A survey," *arXiv preprint arXiv:2509.07996*, 2025.

[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.

[7] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim, "Text-to-image diffusion models in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[8] C. Jia, M. Luo, Z. Dang, G. Dai, X. Chang, M. Wang, and J. Wang, "Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 2480–2488, 2024.

[9] Z. Zhan, D. Chen, J.-P. Mei, Z. Zhao, J. Chen, C. Chen, S. Lyu, and C. Wang, "Conditional image synthesis with diffusion models: A survey," *arXiv preprint arXiv:2409.19365*, 2024.

[10] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9939–9948, 2021.

[11] V. Zyrianov, X. Zhu, and S. Wang, "Learning to generate realistic lidar point clouds," in *European Conference on Computer Vision*, pp. 17–35, Springer, 2022.

[12] Q. Hu, Z. Zhang, and W. Hu, "Rangeldm: Fast realistic lidar point cloud generation," in *European Conference on Computer Vision*, pp. 115–135, Springer, 2024.

[13] Y. Wu, K. Zhang, J. Qian, J. Xie, and J. Yang, "Text2lidar: Text-guided lidar point cloud generation via equirectangular transformer," in *European Conference on Computer Vision*, pp. 291–310, Springer, 2024.

[14] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2147–2156, 2016.

[15] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[16] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21875–21911, 2024.

[17] H.-Y. Tseng, Q. Li, C. Kim, S. Alsisan, J.-B. Huang, and J. Kopf, "Consistent view synthesis with pose-guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16773–16783, 2023.

[18] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[19] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*, pp. 8162–8171, PMLR, 2021.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[21] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

[22] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

[23] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.

[24] L. Zhou, Y. Du, and J. Wu, "3d shape generation and completion through point-voxel diffusion," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5826–5835, 2021.

[25] H. Ran, V. Guizilini, and Y. Wang, "Towards realistic scene generation with lidar diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14738–14748, 2024.

[26] K. Nakashima and R. Kurazume, "Lidar data synthesis with denoising diffusion probabilistic models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14724–14731, IEEE, 2024.

[27] Y. Xiong, W.-C. Ma, J. Wang, and R. Urtasun, "Ultralidar: Learning compact representations for lidar completion and generation," *arXiv preprint arXiv:2311.01448*, 2023.

[28] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[29] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "Rangenet++: Fast and accurate lidar semantic segmentation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4213–4220, 2019.

[30] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," *arXiv preprint arXiv:2202.02703*, 2022.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012.

[33] H. Lin, Z. Guo, Y. Zhang, S. Niu, Y. Li, R. Zhang, S. Cui, and Z. Li, "Drivegen: Generalized and robust 3d detection in driving via controllable text-to-image diffusion generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27497–27507, 2025.

[34] L. Kong, Y. Liu, X. Li, R. Chen, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Robo3d: Towards robust and reliable 3d perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19994–20006, 2023.

[35] Y. Liu, L. Kong, W. Yang, X. Li, A. Liang, R. Chen, B. Fei, and T. Liu, "La la lidar: Large-scale layout generation from lidar data," *arXiv preprint arXiv:2508.03691*, 2025.

[36] A. Liang, Y. Liu, Y. Yang, D. Lu, L. Li, L. Kong, H. Zhao, and W. T. Ooi, "Lidarcrafter: Dynamic 4d world modeling from lidar sequences," *arXiv preprint arXiv:2508.03692*, 2025.

[37] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *European conference on computer vision*, pp. 685–702, Springer, 2020.

[38] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[39] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[40] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[41] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3075–3084, 2019.