La La LiDAR: Large-Scale Layout Generation from LiDAR Data

Youquan Liu[1], Lingdong Kong[2], Weidong Yang[1*], Xin Li[3], Alan Liang[2], Runnan Chen[4*], Ben Fei[5*], Tongliang Liu[4*]

[1]Fudan University
[2]National University of Singapore
[3]Shanghai AI Laboratory
[4]University of Sydney
[5]The Chinese University of Hong Kong

Abstract

Controllable generation of realistic LiDAR scenes is crucial for applications such as autonomous driving and robotics. While recent diffusion-based models achieve high-fidelity LiDAR generation, they lack explicit control over foreground objects and spatial relationships, limiting their usefulness for scenario simulation and safety validation. To address these limitations, we propose **La**rge-scale **La**yout-guided **LiDAR** generation model ("***La La LiDAR***"), a novel layout-guided generative framework that introduces semantic-enhanced scene graph diffusion with relation-aware contextual conditioning for structured LiDAR layout generation, followed by foreground-aware control injection for complete scene generation. This enables customizable control over object placement while ensuring spatial and semantic consistency. To support our structured LiDAR generation, we introduce Waymo-SG and nuScenes-SG, two large-scale LiDAR scene graph datasets, along with new evaluation metrics for layout synthesis. Extensive experiments demonstrate that La La LiDAR achieves state-of-the-art performance in both LiDAR generation and downstream perception tasks, establishing a new benchmark for controllable 3D scene generation.

1 Introduction

Generating 3D scenes has emerged as a critical technology for a wide spectrum of real-world applications, including autonomous driving, AR/VR, and robotics (Muhammad et al. 2022; Bian et al. 2025; Lee et al. 2024). In particular, for autonomous driving systems, the quality, diversity, and controllability of 3D training data directly impact perception capabilities and generalization performance (Wilson et al. 2022; Zheng et al. 2024; Xie et al. 2025). Traditional data collection methods, however, face substantial limitations in terms of cost, scalability, and coverage of rare scenarios, making synthetic data generation an increasingly attractive alternative. Recent advancements in generative models, especially diffusion-based approaches (Nichol and Dhariwal 2021; Rombach et al. 2022), have demonstrated remarkable success in high-fidelity image synthesis. This progress has naturally extended to 3D representations, yielding promising results for LiDAR point cloud generation in driving scenarios (Hu, Zhang, and Hu 2024; Nunes et al. 2024).

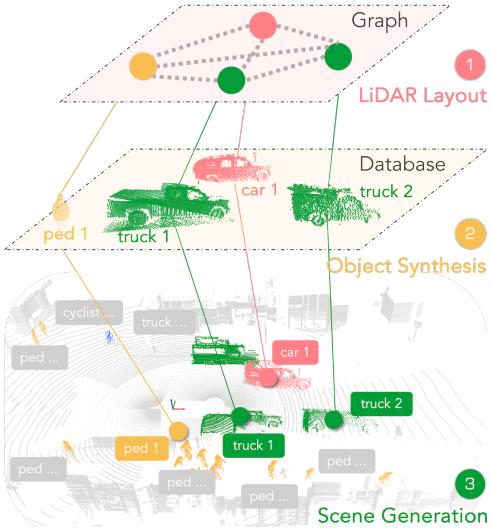


Figure 1: Motivation of customizable LiDAR scene generation from "La La LiDAR". Our framework consists of three key stages: 1) LiDAR layout generation using scene graphs, where nodes represent objects and edges capture their spatial relationships; 2) foreground point cloud synthesis, either by retrieving from a database or by generating conditioned on layout parameters; and 3) foreground-conditioned scene generation, where synthesized foreground serve as conditioning to generate the complete scene with realistic environmental context. This hierarchical approach enables fine-grained control over foreground object placement while maintaining overall scene coherence.

Despite notable advances in unconditional LiDAR scene generation, current approaches suffer from a critical limitation: the lack of customizable control over foreground object composition and arrangement. This deficiency is particularly significant in autonomous driving scenarios, where foreground objects such as vehicles, pedestrians, and other traffic participants constitute only a small fraction (approximately 6.29%, measured on nuScenes) of the total point cloud. Although these foreground elements are crucial for downstream perception tasks, existing methods process foreground and background elements uniformly, offer-

*Corresponding author

ing minimal control over object placement and inter-object spatial relationships. This uniform treatment limits the applicability of these generative models for driving simulation and long-tail scenario synthesis.

To address these limitations, we propose **La La LiDAR** (Large-scale Layout-guided LiDAR) as shown in Figure 1, a novel framework for fine-grained controllable LiDAR scene generation that explicitly models spatial relationships between foreground objects through structured scene graph representations. Our approach comprises three main steps: **1)** Structured layout generation: synthesizing LiDAR layouts via a scene graph-based diffusion model to ensure spatial and semantic coherence. **2)** Foreground point cloud synthesis: retrieving or generating object-level point clouds conditioned on the layout. **3)** Foreground-conditioned scene generation: leveraging the synthesized foreground as control signals to generate complete LiDAR scene with realistic environmental context. By integrating graph-based spatial reasoning and diffusion-based generative modeling, our method enables fine-grained customization of scene composition while preserving the realism of large-scale LiDAR environments. The contributions of our work include:

- We introduce **La La LiDAR**, a comprehensive LiDAR layout generation framework that leverages scene graphs to capture semantic and spatial relationships among foreground objects. This enables customizable scene composition while maintaining physical plausibility.
- We propose a novel Foreground-aware Control Injector (FCI) that effectively bridges layout information with the scene generation process, ensuring accurate representation and spatial coherence of foreground elements within the broader environmental context.
- We construct two large-scale LiDAR scene graph datasets, **Waymo-SG** and **nuScenes-SG**, along with specialized evaluation metrics tailored for LiDAR layout generation, establishing new benchmarks for future research in LiDAR layout synthesis.
- Our analyses prove superior performance of our approach in LiDAR layout generation, LiDAR scene generation, and multiple downstream perception tasks, including LiDAR semantic segmentation, 3D object detection, and scene completion.

Extensive experiments demonstrate that La La LiDAR not only produces high-fidelity scene but also provides unprecedented control over scene composition. By explicitly modeling foreground objects and their relationships, our method addresses a critical need in the development and evaluation of autonomous driving systems, enabling the generation of diverse, realistic, and controllable driving scenarios.

2 Related Work

LiDAR Scene Generation. LiDARGen (Zyrianov, Zhu, and Wang 2022) pioneered diffusion-based LiDAR generation using range and reflectance data. UltraLiDAR (Xiong et al. 2023) integrates voxelized LiDAR point cloud with a VQ-VAE framework (Van Den Oord, Vinyals et al. 2017), thereby enabling efficient LiDAR data generation and completion. R2DM (Nakashima and Kurazume 2024) improved generation quality with advanced denoising networks, while LiDM (Ran, Guizilini, and Wang 2024) focused on preserving geometric structures. RangeLDM (Hu, Zhang, and Hu 2024) emphasized real-time generation, and Text2LiDAR (Wu et al. 2024) introduced text-guided synthesis for semantic control. However, existing methods treat foreground and background uniformly, limiting control over object placement and spatial relationships. To address this, we propose a layout-guided LiDAR generation framework for more controllable and diverse LiDAR scene generation.

3D Generation from Scene Layouts. While extensively studied in indoor environments, scene layout-based generation remains underexplored for outdoor settings. Methods like Graph-to-3D (Dhamo et al. 2021) and CommonScenes (Zhai et al. 2023) utilize scene graphs for spatially coherent room generation, while EchoScene (Zhai et al. 2024) further improves layout consistency through information echo mechanisms. We introduce the first scene graph-based approach for outdoor LiDAR generation, which explicitly models foreground objects and their spatial relationships, and proposes tailored evaluation metrics.

Downstream Applications. LiDAR-based perception plays a crucial role in autonomous driving (Kong et al. 2023a; Liu et al. 2024, 2023; Li et al. 2023). For LiDAR semantic segmentation, methods like SPVCNN (Tang et al. 2020) leverages sparse convolutions, while semi-supervised approaches such as LaserMix (Kong et al. 2023b) improves data efficiency. 3D object detection models like SECOND (Yan, Mao, and Li 2018) localize traffic participants using voxelized representations, directly impacting navigation safety. LiDAR completion tackles occlusion, with diffusion models (Zhao et al. 2025; Martyniuk et al. 2025) achieving superior reconstruction via learned priors. Our work enhances these applications by generating high-quality LiDAR with controllable foreground elements, improving robustness in segmentation, detection, and completion.

3 Methodology

We propose a layout-guided LiDAR generation framework comprising two stages: (i) layout generation and foreground synthesis, (ii) foreground-aware scene generation. This design enables customized control over foreground objects while ensuring spatial coherence, addressing the limitations of existing LiDAR scene generation approaches.

3.1 Preliminaries

LiDAR Representation. A LiDAR point cloud is defined as $\mathcal{P} = \{(\mathbf{p}^i, \mathbf{e}^i) \mid i = 1, ..., N\}$, where each point $\mathbf{p}^i \in \mathbb{R}^3$ represents the 3D coordinates (p_x^i, p_y^i, p_z^i) and $\mathbf{e}^i \in \mathbb{R}^L$ denotes auxiliary attributes such as intensity. Following prior works (Zyrianov, Zhu, and Wang 2022; Wu et al. 2024), we adopt a spherical projection (Milioto et al. 2019a,b) to convert $\mathcal{P}$ into a structured range image $X \in \mathbb{R}^{H \times W \times 2}$ for efficient processing. The projection process $\Pi : \mathbb{R}^3 \mapsto \mathbb{R}^2$ is

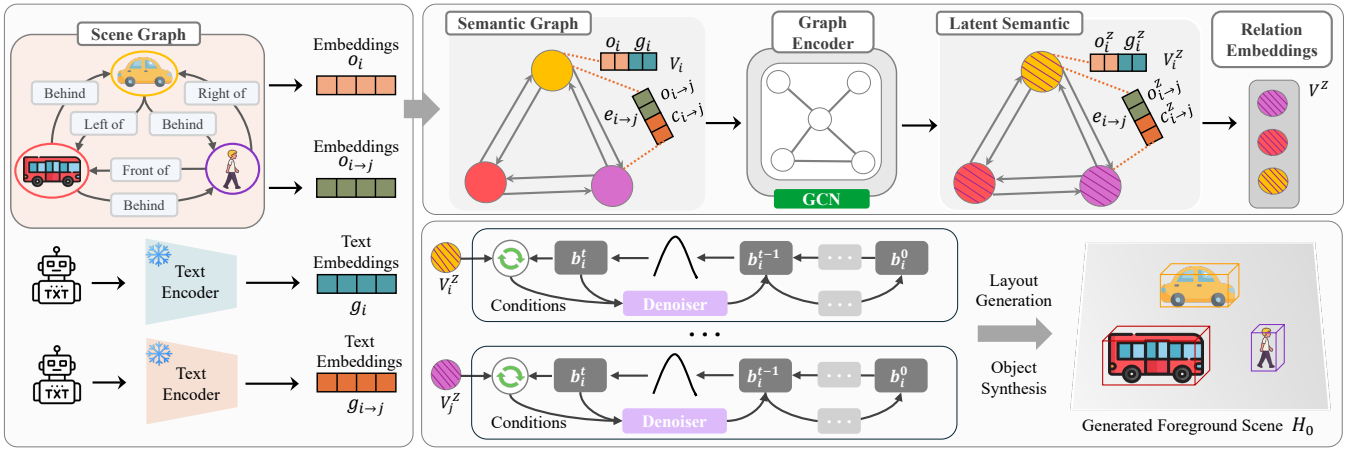


Figure 2: The proposed LiDAR point cloud layout generation framework. Our approach begins with scene graph construction, establishing both node embeddings (o_i) and edge embeddings ($o_{i\to j}$) to capture spatial relationships. These are enhanced with semantic features from a CLIP text encoder (g_i, $g_{i\to j}$), creating a comprehensive semantic graph. Graph Encoder then processes this information to produce a latent semantic graph with enriched node representations (V_i^Z). During the diffusion process, layout states (b_i^t) are iteratively refined through a denoising network that incorporates time-dependent contextual conditioning ($\mathcal{C}_t$), which dynamically aggregates graph features at each timestep. This ensures consistent spatial relationships throughout the denoising process. The final stage synthesizes and places appropriate foreground points according to the generated layout.

defined as follows:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \frac{\arctan(p_y^i, p_x^i)}{\pi}\right] W \\ \left[1 - \frac{\arcsin(p_z^i/d) + f_{\text{up}}}{f}\right] H \end{pmatrix}, \quad (1)$$

where $d = \|\mathbf{p}^i\|_2$ is the depth, $f = f_{\text{up}} + f_{\text{down}}$ is the vertical field-of-view, and (H, W) denote the vertical and horizontal resolutions. Each pixel in X stores the depth and intensity.

Conditional Diffusion Models. We adopt the denoising diffusion probabilistic model (DDPM) (Nichol and Dhariwal 2021) conditioned on structured inputs. The model learns to predict the Gaussian noise ϵ added to clean data $\mathbf{x}_0$ through a noise prediction network ϵ_θ. The training objective:

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, c)\|_2^2 \right], \quad (2)$$

where $\mathbf{x}_t$ is the noisy sample at timestep t, and c denotes a conditioning signal such as semantic layout or foreground. This framework enables controllable generation, which we exploit in both layout and scene synthesis stages.

3.2 LiDAR Layout Generation

Achieving realistic and controllable LiDAR scene generation requires explicitly modeling the spatial arrangement of foreground objects. To this end, we introduce a structured layout generation framework in Figure 2, which leverages scene graphs to represent semantic and spatial relations among objects and guides the synthesis of object layouts.

LiDAR Scene Graph Construction. We represent each LiDAR scene as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $v_i \in \mathcal{V}$ denote foreground objects, and edges $e_{i\to j} \in \mathcal{E}$ capture pairwise spatial or semantic relations. We embed v_i and $e_{i\to j}$ in the scene graph to derive node embeddings o_i and edge embeddings $o_{i\to j}$. Each node and edge is labeled with class attributes, i.e., $c_i^{\text{node}} \in \mathcal{C}^{\text{node}}$ and $c_{i\to j}^{\text{edge}} \in \mathcal{C}^{\text{edge}}$, respectively. Each object layout b_i is parameterized by a 3D bounding box in the ego-vehicle frame, including (x, y, z) position, (l, h, w) size, and yaw θ. All parameters are normalized, and θ is encoded using sine-cosine to preserve angular continuity. Due to the lack of existing LiDAR scene graph benchmarks, we construct two datasets: **Waymo-SG** and **nuScenes-SG**, derived from Waymo Open and nuScenes datasets, respectively. These graphs encode nine relation types, including spatial (e.g., *front of*, *left of*) and comparative (e.g., *bigger than*, *taller than*) relations.

Graph-Based Semantic Encoding. To model structured spatial priors, we represent the input scene as a semantic graph and process it with a triplet Graph Convolutional Network (GCN). Following (Johnson, Gupta, and Fei-Fei 2018; Zhai et al. 2023), our GCN iteratively updates node and edge features by aggregating contextual messages. At each layer k, node v_i and edge $e_{i\to j}$ are updated as:

$$\begin{aligned} (\alpha_{v_i}^{(k)}, \beta_{e_{i\to j}}^{(k+1)}, \alpha_{v_j}^{(k)}) &= \text{MLP}_1(\beta_{v_i}^{(k)}, \beta_{e_{i\to j}}^{(k)}, \beta_{v_j}^{(k)}), \\ \beta_{v_i}^{(k+1)} &= \alpha_{v_i}^{(k)} + \text{MLP}_2\left(\text{Avg}\left(\alpha_{v_j}^{(k)} \mid v_j \in N_{\mathcal{G}}(v_i)\right)\right), \end{aligned} \quad (3)$$

where MLP_1, MLP_2 are multi-layer perceptrons, and $N_{\mathcal{G}}(v_i)$ denotes neighbors of v_i in the scene graph.

To enhance semantic alignment, we leverage CLIP (Radford et al. 2021) encoder E_{clip} to encode nodes and edges in a unified language-vision space. Given an object class c_i^{node}, we compute $g_i = E_{\text{clip}}(c_i^{\text{node}})$. Relations are embedded via templated prompts (e.g., "Car [left of] Pedestrian"), yielding $g_{i\to j} = E_{\text{clip}}(c_i^{\text{node}}, c_{i\to j}^{\text{edge}}, c_j^{\text{node}})$. These features are concatenated with learnable embeddings to form enriched node descriptors: $V_i = \text{Concat}(g_i, o_i)$. The initial GCN inputs are set as $(\beta_{v_i}^{(0)}, \beta_{e_{i\to j}}^{(0)}, \beta_{v_j}^{(0)}) = (V_i, e_{i\to j}, V_j)$. After K

GCN layers, we obtain latent node embeddings $\{V_i^z\}$ that encode both semantic class and relational structure. These embeddings serve as the foundation for the subsequent layout diffusion process.

Layout Diffusion. We formulate layout generation as a conditional diffusion process, where each foreground object node v_i is associated with a time-dependent noisy layout state $b_i^t \in \mathbb{R}^8$. To enable relationally grounded denoising, we construct a node-wise input embedding by concatenating its semantic feature V_i^z, the temporal encoding $\pi(t)$, and the noisy layout b_i^t:

$$\tilde{V}_i^t = \text{Concat}(V_i^z, \pi(t), b_i^t) \, . \tag{4}$$

These features are processed by a GCN over the scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, producing a time-varying contextual embedding $\mathcal{C}_t$ that aggregates relational priors across the graph. Given the diffusion schedule coefficients $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, reverse step updates each node's layout as:

$$b_i^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(b_i^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(b_i^t, t, \mathcal{C}_t) \right) + \sigma_t \mathbf{z} \, , \tag{5}$$

where $\mathbf{z} \sim \mathcal{N}(0, I)$ for $t > 1$, and $\mathbf{z} = 0$ otherwise. The denoising network ϵ_θ is implemented as a cross-attention transformer, shared across nodes and conditioned on $\mathcal{C}_t$. Starting from Gaussian noise $b_i^T \sim \mathcal{N}(0, I)$, the model progressively refines object layouts over T steps. By incorporating dynamically aggregated graph context, our model avoids node-level isolation (Zhai et al. 2024) and enforces globally consistent spatial configurations during generation.

Training Objectives. To promote realistic and physically consistent LiDAR layouts, we optimize a combination of geometric and diffusion-based objectives. Let $\{\hat{b}_i\}_{i=1}^M$ denote predicted layouts and $\{b_i\}_{i=1}^M$ the ground truth. We first minimize spatial collisions via a pairwise 3D IoU-based penalty:

$$\mathcal{L}_{\text{collision}} = \frac{1}{M} \sum_{i \neq j} \max(\text{IoU}(\hat{b}_i, \hat{b}_j) - \delta, 0) \, , \tag{6}$$

where δ is a small tolerance threshold. To enforce alignment with ground truth geometry, we introduce an IoU-based loss:

$$\mathcal{L}_{\text{IoU}} = \frac{1}{M} \sum_{i} (1 - \text{IoU}(\hat{b}_i, b_i)) \, . \tag{7}$$

Additionally, we apply a standard diffusion reconstruction loss that supervises the denoising network by predicting the injected noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{b_i, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\epsilon - \epsilon_\theta(b_i^t, t, \mathcal{C}_t)\|_2^2 \right] \, . \tag{8}$$

The final objective is a weighted sum of the three terms:

$$\mathcal{L}_{\text{layout}} = \lambda_1 \mathcal{L}_{\text{collision}} + \lambda_2 \mathcal{L}_{\text{IoU}} + \lambda_3 \mathcal{L}_{\text{diff}} \, , \tag{9}$$

where λ_1, λ_2, and λ_3 are balancing coefficients.

Object Synthesis. After generating the scene layouts, we populate them with foreground objects via either retrieval or generation of their LiDAR point clouds. For retrieval, an object database built from nuScenes-SG and Waymo-SG is queried to match each layout box with a semantically aligned instance of similar shape. To enhance diversity and address rare categories, we additionally train a DiT-3D based (Mo et al. 2023) object generator conditioned on category and layout parameters. Both pathways ensure spatial consistency and semantic coherence in object placement.

3.3 3D Scene Generation

To enable customizable LiDAR scene generation, we generate background content conditioned on the foreground point cloud derived from the layout. In contrast to prior LiDAR generation methods that treat all points uniformly and often underrepresent sparse yet critical foreground regions, our framework ensures both semantic controllability and spatial consistency across the entire scene.

Conditional Scene Generation. To condition the denoising process on structured foreground geometry, we introduce the Foreground-aware Control Injector (FCI), shown in Figure 3. Given the foreground point cloud H_0, we extract multi-scale features $\{\hat{H}_i\}_{i=1}^l$ using ResBlocks. For each injection layer, a corresponding $\hat{H}_i$ is transformed into channel-wise scale and shift parameters to modulate intermediate features. To suppress invalid regions from sparse foreground input, a binary mask $X_m \in \{0,1\}^{h \times w}$ is applied before nonlinearity at each resolution.

In addition, we introduce a gating mechanism to adaptively weight feature contributions. Specifically, each $\hat{H}_i$ is processed through a convolutional branch with dimension reduction, SiLU activation, and expansion, yielding a spatial attention map $\omega \in \mathbb{R}^{1 \times h \times w}$. The denoiser feature $X_f^t \in \mathbb{R}^{C \times h \times w}$ is then modulated as:

$$X_p^t = X_f^t \cdot (1 + \text{scale} \cdot \omega) + \text{shift} \cdot \omega \, . \tag{10}$$

We inject this conditioning module at multiple scales to maintain semantic awareness and spatial alignment throughout the generative hierarchy.

Training Objectives. We adopt a noise reconstruction objective similar to layout generation. The primary loss minimizes the mean squared error between the predicted noise $\varepsilon_\theta(X_t, t, H_0)$ and the ground-truth noise ϵ at timestep t:

$$\mathcal{L}_{\text{scene}} = \mathbb{E}_{X, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\epsilon - \epsilon_\theta(X^t, t, H_0)\|_2^2 \right] \, . \tag{11}$$

To emphasize fidelity in foreground regions, we introduce an auxiliary foreground loss $\mathcal{L}_{\text{fore}}$, which restricts the denoising error within the masked foreground area X_m:

$$\mathcal{L}_{\text{fore}} = \mathbb{E}_{X, \epsilon \sim \mathcal{N}(0,I), t} \left[\|(\epsilon - \epsilon_\theta(X^t, t, H_0)) \odot X_m\|_2^2 \right] \, . \tag{12}$$

The overall training loss is a weighted sum of the two:

$$\mathcal{L}_{\text{cond}} = \lambda_4 \mathcal{L}_{\text{scene}} + \lambda_5 \mathcal{L}_{\text{fore}} \, , \tag{13}$$

where λ_4 and λ_5 weight each term.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on two large-scale autonomous driving datasets: nuScenes (Caesar et al. 2020) and Waymo Open (Sun et al. 2020). To support layout-conditioned generation, we construct two scene graph datasets, nuScenes-SG and Waymo-SG, by extracting spatial and semantic relationships among foreground objects. Additional details on scene graph construction are provided in the Appendix.

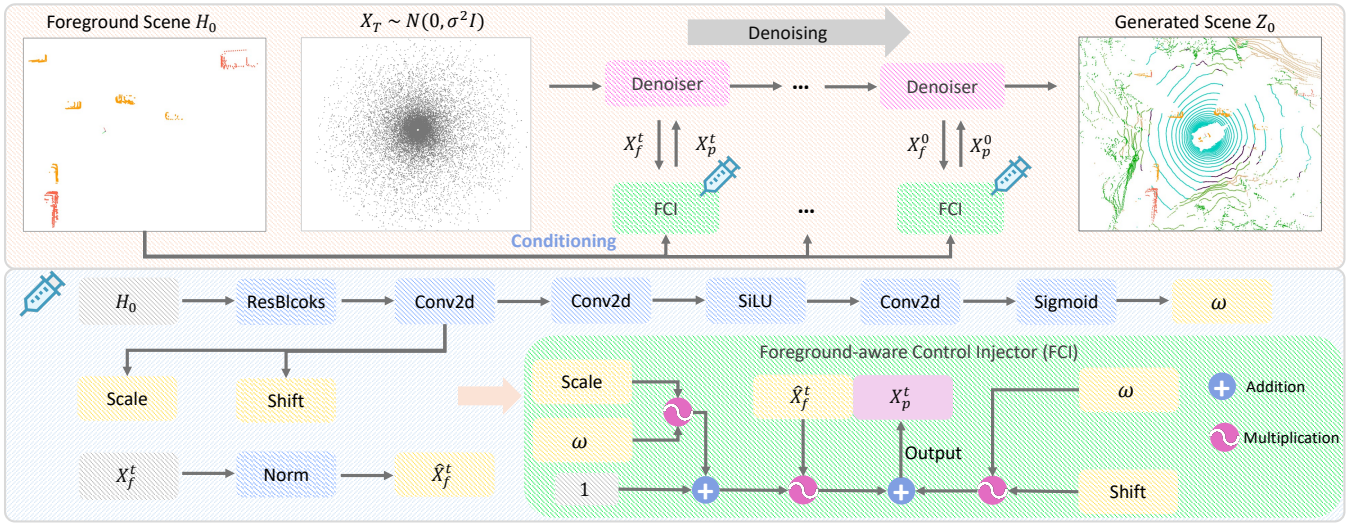


Figure 3: Architecture of our foreground-aware LiDAR scene generation framework. *Upper Part*: The diffusion-based generation process, where initial Gaussian noise $X_T \sim \mathcal{N}(0, \sigma^2 I)$ is progressively denoised to generate the final scene Z_0, conditioned on a foreground input H_0 via our FCI. *Lower Part*: The FCI mechanism extracts features from H_0 and transforms them into adaptive scale and shift parameters. These modulate the intermediate features X_f^t in the denoising network through channel-wise gating with attention weights ω, resulting in refined features X_p^t that preserve object details. This design ensures spatial coherence and semantic consistency in the generated scene Z_0.

Method	Venue	nuScenes RAE ↑	nuScenes RAD ↑	nuScenes CR ↓	nuScenes $\text{Prec}_{0.3}$ ↑	nuScenes $\text{Prec}_{0.5}$ ↑	nuScenes IoU ↑	Waymo Open RAE ↑	Waymo Open RAD ↑	Waymo Open CR ↓	Waymo Open $\text{Prec}_{0.3}$ ↑	Waymo Open $\text{Prec}_{0.5}$ ↑	Waymo Open IoU ↑
Graph-to-Box	ICCV'21	0.80	0.37	0.13	4.69	4.35	17.80	0.74	0.18	0.32	3.69	3.39	11.35
CommonScenes	NeurIPS'23	0.83	0.45	0.13	4.85	4.49	18.71	0.78	0.24	0.30	4.44	4.10	13.73
EchoScene	ECCV'24	0.91	0.65	0.07	5.48	5.12	26.70	**0.91**	0.55	0.17	5.33	5.02	15.67
La La LiDAR	**Ours**	**0.92**	**0.68**	**0.06**	**6.59**	**6.15**	**28.14**	**0.91**	**0.57**	**0.14**	**6.39**	**6.08**	**18.64**

Table 1: Comparisons of state-of-the-art Layout Generation approaches on the val sets of nuScenes and Waymo Open. Metrics with ↓ indicate lower is better. All $\text{Prec}_{0.3}$, $\text{Prec}_{0.5}$ and IoU scores are given in percentage (%).

Method	Venue	FRD↓	FPD↓	JSD↓	MMD↓
LiDARGen	ECCV'22	549.2	22.8	0.04	0.8
R2DM	ICRA'24	253.8	14.4	**0.03**	**0.5**
LiDM	CVPR'24	-	30.8	0.07	3.9
Text2LiDAR	ECCV'24	953.2	147.5	0.09	12.5
La La LiDAR	**Ours**	**211.0**	**9.8**	0.04	0.7

Table 2: Comparisons of LiDAR Scene Generation methods on the nuScenes dataset. The MMD metric is in 10^{-4}.

Evaluation Metrics. We evaluate LiDAR layout generation using four metrics: Relationship Accuracy Easy (RAE), Relationship Accuracy Difficult (RAD), Collision Rate (CR), and Intersection over Union (IoU). For LiDAR scene generation, we report Fréchet Range Distance (FRD), Fréchet Point Cloud Distance (FPD), Jensen-Shannon Divergence (JSD), and Maximum Mean Discrepancy (MMD). For downstream tasks, we use mean IoU (mIoU) for semantic segmentation, mean Average Precision (mAP) and nuScenes Detection Score (NDS) for 3D object detection, and Mean Absolute Error (MAE) for scene completion.

4.2 Comparative Study

LiDAR Layout Generation. Table 1 shows that our method outperforms previous methods across all metrics on both datasets. On nuScenes, we achieve higher precision (6.59% and 6.15%) and IoU (28.14%), with better relationship accuracy (RAE: 0.92, RAD: 0.68) and lower collision rate (0.06). Similar improvements are observed on Waymo Open. These results demonstrate the effectiveness of our approach in generating spatially coherent foreground layouts.

LiDAR Scene Generation. Our method achieves state-of-the-art performance on LiDAR scene generation, as shown in Table 2. La La LiDAR fulfills the lowest FPD (9.8) and FRD (211.0), outperforming LiDARGen (22.8, 549.2) and R2DM (14.4, 253.8). These results indicate that our model generates LiDAR samples with both high geometric fidelity and semantic alignment. Figure 4 presents qualitative comparisons against LiDARGen and R2DM. LiDARGen introduces artifacts in the background and fails to maintain structural consistency. R2DM exhibits a smoother global layout but lacks foreground sharpness and coherent fusion with surrounding regions. In contrast, our model produces scenes with detailed foreground geometry and seam-

Base	Method	Venue	nuScenes 1%	10%	20%	50%
MinkU.	*Sup.-only*	-	58.3	71.0	73.0	75.1
	MeanTeacher	NeurIPS'17	60.1	71.7	73.4	75.2
	LaserMix	CVPR'23	62.8	73.6	74.8	76.1
	R2DM	ICRA'24	64.1	73.0	74.3	75.9
	La La LiDAR	**Ours**	**65.1**	**73.8**	**75.4**	**76.2**
SPVCNN	*Sup.-only*	-	57.9	71.7	73.0	74.6
	MeanTeacher	NeurIPS'17	59.4	72.5	73.1	74.7
	LaserMix	CVPR'23	63.2	**74.1**	74.6	75.8
	R2DM	ICRA'24	64.6	72.7	74.2	75.4
	La La LiDAR	**Ours**	**65.4**	74.0	**75.0**	**76.3**

Table 3: Downstream application of La La LiDAR for the 3D Semantic Segmentation task on the val set of nuScenes.

Base	Method	1% mAP	1% NDS	5% mAP	5% NDS	10% mAP	10% NDS	20% mAP	20% NDS
Center.	*Sup.-only*	24.7	31.2	36.7	40.1	40.9	43.0	40.7	42.9
	R2DM	26.4	32.1	36.9	41.2	41.0	43.8	41.9	45.0
	Text2LiDAR	26.0	31.8	36.8	40.7	41.2	43.9	41.4	44.9
	La La LiDAR	**27.0**	**32.6**	**37.0**	**41.6**	**41.4**	**44.4**	**42.5**	**45.6**
SEC.	*Sup.-only*	0.8	19.4	29.2	36.5	34.5	41.2	34.3	41.3
	R2DM	13.1	21.2	32.7	39.4	36.2	42.1	39.7	44.5
	Text2LiDAR	15.0	25.4	31.3	38.5	34.6	41.4	37.1	43.3
	La La LiDAR	**15.4**	**25.8**	**33.1**	**40.0**	**36.6**	**42.6**	**39.9**	**44.9**

Table 4: Downstream application of La La LiDAR for the 3D Object Detection task on the val set of nuScenes.

less background integration. This visual fidelity benefits from our FCI mechanism, which ensures accurate conditioning and structure-aware denoising. These qualitative and quantitative results together validate the effectiveness of our method in generating high-quality, controllable scenes.

LiDAR Semantic Segmentation. Moreover, our method could be tamed as a generative data augmentation strategy for LiDAR segmentation, where generated samples (pseudo-labeled using a pre-trained SPVCNN) are combined with limited real data during training. As shown in Table 3, our method consistently improves performance, outperforming both R2DM and the semi-supervised method LaserMix. These results demonstrate the effectiveness of our generated scenes in enhancing label efficiency for 3D segmentation.

3D Object Detection. We further evaluate La La LiDAR for detection by augmenting training data with generated samples. As shown in Table 4, our method consistently improves performance across various data regimes. These findings validate the effectiveness of our approach in enhancing detector training when real annotations are scarce.

Scene Completion. Furthermore, we evaluate our framework on LiDAR completion using a Repaint (Lugmayr et al. 2022) where every fourth beam is preserved. As shown in Table 5, we outperform both interpolation and diffusion-based approaches across all metrics, demonstrating superior structural and semantic coherence during completion.

4.3 Ablation Study

Analysis of Key Design Choices for Layout Generation. We conduct ablation studies in Table 6 to assess three key

Base	Method	MAE ↓ (Range)	MAE ↓ (Reflectance)	IoU ↑ (Semantics)
Inter.	Nearest Neighbor	4.00	8.16	25.11
	Bi-Linear	4.27	8.59	31.27
	Bi-Cubic	4.59	9.10	25.23
Diff.	R2DM	2.64	5.63	64.44
	La La LiDAR (Ours)	**2.45**	**5.10**	**66.90**

Table 5: Downstream application of La La LiDAR for the LiDAR Completion task on the nuScenes. Inter.: Interpolation methods. Diff.: Diffusion-based methods.

#	Configuration	RAE↑	RAD↑	CR ↓	$\text{Prec}_{0.3}$↑	$\text{Prec}_{0.5}$↑	IoU ↑
a	Ours *w/o* SG	0.91	0.59	0.08	5.01	4.76	23.64
b	Ours *w/* Concat condition	**0.92**	0.63	0.07	5.15	4.82	25.89
c	Ours *w/o* OI	**0.92**	0.62	0.08	5.96	5.50	26.84
d	**The Full Framework**	**0.92**	**0.68**	**0.06**	**6.59**	**6.15**	**28.14**

Table 6: Ablation study of the outdoor LiDAR layout generation on the nuScenes dataset. SG is the semantic graph, and OI refers to the combination of $\mathcal{L}_{\text{collosion}}$ and $\mathcal{L}_{\text{IoU}}$.

#	Configuration	FRD↓	FPD↓	JSD↓	MMD↓
a	Baseline	253.79	14.41	**0.034**	**0.54**
b	Box layout condition	244.60	14.22	0.051	1.24
c	Ours *w/o* FM	301.34	19.45	0.048	1.66
d	Ours *w/* Add condition	235.37	**9.38**	0.047	1.38
e	Ours *w/* Concat condition	232.15	9.65	0.047	1.30
f	Ours *w/* CA condition	461.85	27.47	0.091	3.77
g	**The Full Framework**	**211.03**	9.82	0.039	0.74

Table 7: Ablation study of our LiDAR scene generation framework on the nuScenes dataset. FM refers to foreground mask and CA denotes cross-attention.

designs of our layout generation framework: semantic graph encoding, conditioning mechanism, and geometric consistency losses. First, excluding semantic features (row a vs. d) markedly reduces relationship accuracy (RAD: 0.68 → 0.59) and spatial precision (IoU: 28.14% → 23.64%), highlighting the contribution of scene-level semantics. Second, replacing cross-attention with naive concatenation (row b vs. d) leads to degraded RAD (0.68 → 0.63) and object alignment ($\text{Prec}_{0.5}$: 6.15% → 4.82%), indicating that attention mechanisms better capture relational structure. Finally, ablating our geometric loss terms (row c vs. d) increases the collision rate and further reduces IoU, confirming their effectiveness in promoting physically consistent layouts.

Ablation on Foreground Mask and Conditioning Strategy. We ablate two key components of our scene generation framework: the foreground validity mask X_m and the FCI mechanism. Removing X_m (row c in Table 7) leads to a notable degradation across all metrics (e.g., FPD: 9.8 → 19.5), confirming the necessity of suppressing invalid sparse regions during feature modulation. To evalute conditioning strategies, we compare addition (row d), concatenation (row e), and cross-attention alone (row f). While both addition and concatenation outperform the baseline, cross-attention performs poorly (FPD: 27.5), likely due to unstable modu-

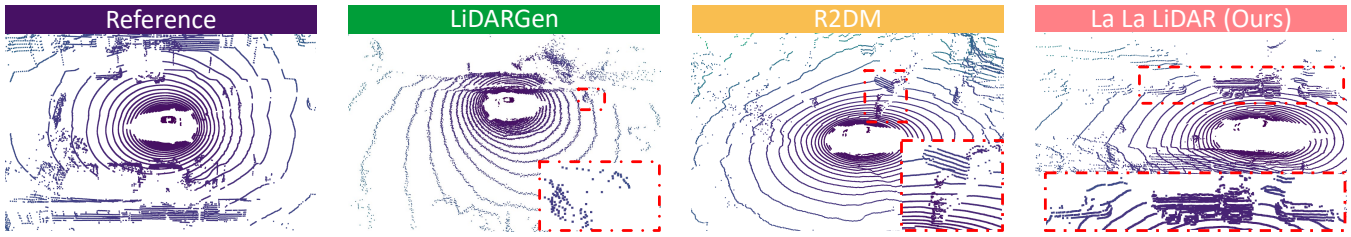
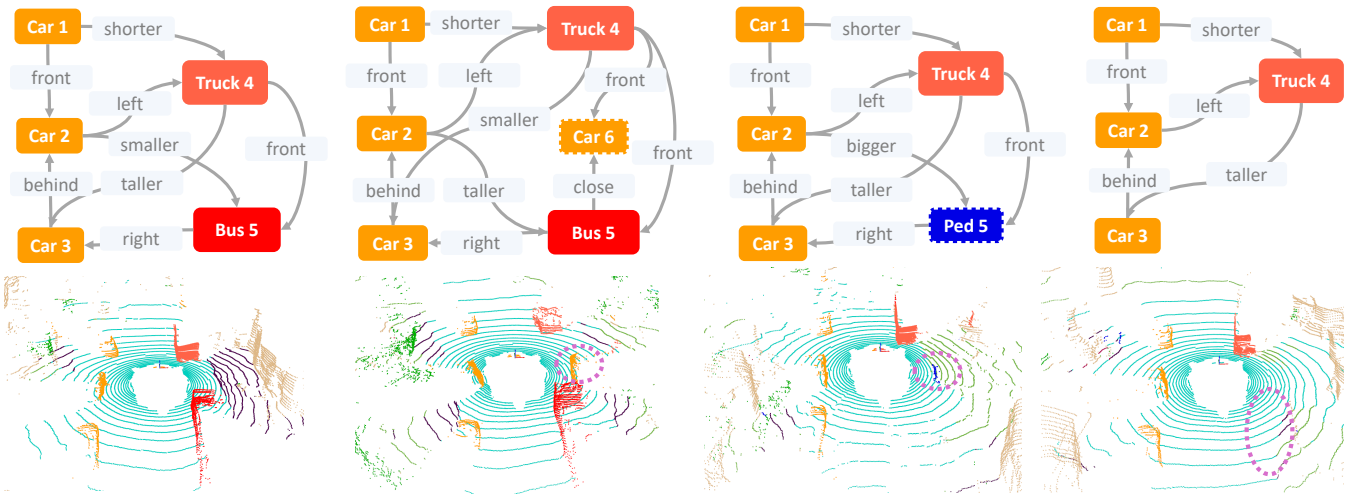


Figure 4: Qualitative comparisons of La La LiDAR against state-of-the-art LiDAR scene generation approaches on the nuScenes dataset. From left to right: Reference (ground truth), LiDARGen, R2DM, and our method.



(a) Original (b) Addition (c) Replacement (d) Removal

Figure 5: Controllable scene synthesis via graph-level editing. Top: input scene graphs with object nodes and relational edges; Bottom: generated LiDAR scenes. Edge visualizations are partially omitted for clarity.

lation. Our final model (row g), which integrates X_m and the FCI module, achieves the best overall performance, demonstrating that structured and spatially aligned conditioning is essential for controllable and coherent scene synthesis.

Advantages of Scene Graph Conditioning over Box Layout. To assess the benefit of structured relational priors, we compare our scene graph-based conditioning with a baseline that uses only 3D box layouts (row b in Table 7). Despite providing geometric cues, the box-only representation underperforms our method in both semantic consistency (FRD: 244.6 $\rightarrow$ 211.0) and geometric fidelity (FPD: 14.2 $\rightarrow$ 9.8), highlighting the value of explicit relational structure in guiding scene generation. Beyond quantitative gains, scene graphs offer key advantages: (1) They encode inter-object relationships (e.g., *car behind pedestrian*) that box layout cannot represent, enabling richer spatial reasoning; (2) They support hierarchical control, allowing users to specify both object presence and their relational configuration; (3) They serve as structured, extensible priors that integrate naturally with downstream autonomy modules such as planning or forecasting. These insights highlight the advantages of structured relational priors for controllable LiDAR generation.

Graph-Driven Customization of Scene Generation. Figure 5 illustrates the controllable generation of our framework via scene graph editing. Starting from an initial graph (a), we apply three types of modifications: (b) inserting a new object with relational edges, (c) substituting an object category, and (d) removing an object node. Each operation induces coherent changes in the generated LiDAR scenes, showing the model's ability to respond to structured semantic edits. These results highlight the flexibility of our approach in enabling fine-grained, relation-aware scene manipulation.

5 Conclusion

We propose **La La LiDAR**, a novel layout-guided generative framework for controllable LiDAR scene generation. Our approach enables controllable control over object placement and spatial relationships, addressing the limitations of existing unconditional LiDAR generation methods. To support structured LiDAR generation, we introduce Waymo-SG and nuScenes-SG, two large-scale LiDAR scene graph datasets, along with new evaluation metrics specifically designed for LiDAR layout generation. Extensive experiments demonstrate that our method achieves leading performance in both LiDAR generation and downstream perception tasks, validating its effectiveness for real-world applications.

References

Bian, H.; Kong, L.; Xie, H.; Pan, L.; Qiao, Y.; and Liu, Z. 2025. DynamicCity: Large-Scale 4D Occupancy Generation from Dynamic Scenes. In *International Conference on Learning Representations*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Dhamo, H.; Manhardt, F.; Navab, N.; and Tombari, F. 2021. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *IEEE/CVF International Conference on Computer Vision*, 16352–16361.

Hu, Q.; Zhang, Z.; and Hu, W. 2024. RangeLDM: Fast Realistic LiDAR Point Cloud Generation. In *European Conference on Computer Vision*, 115–135. Springer.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *CVPR*.

Kong, L.; Liu, Y.; Li, X.; Chen, R.; Zhang, W.; Ren, J.; Pan, L.; Chen, K.; and Liu, Z. 2023a. Robo3D: Towards Robust and Reliable 3D Perception against Corruptions. In *IEEE/CVF International Conference on Computer Vision*, 19994–20006.

Kong, L.; Ren, J.; Pan, L.; and Liu, Z. 2023b. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21705–21715.

Lee, J.; Lee, S.; Jo, C.; Im, W.; Seon, J.; and Yoon, S.-E. 2024. Semcity: Semantic scene generation with triplane diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 28337–28347.

Li, X.; Ma, T.; Hou, Y.; Shi, B.; Yang, Y.; Liu, Y.; Wu, X.; Chen, Q.; Li, Y.; Qiao, Y.; et al. 2023. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17524–17534.

Liu, Y.; Chen, R.; Li, X.; Kong, L.; Yang, Y.; Xia, Z.; Bai, Y.; Zhu, X.; Ma, Y.; Li, Y.; Qiao, Y.; and Hou, Y. 2023. UniSeg: A Unified Multi-Modal LiDAR Segmentation Network and the OpenPCSeg Codebase. In *IEEE/CVF International Conference on Computer Vision*, 21662–21673.

Liu, Y.; Kong, L.; Wu, X.; Chen, R.; Li, X.; Pan, L.; Liu, Z.; and Ma, Y. 2024. Multi-Space Alignments Towards Universal LiDAR Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14648–14661.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11461–11471.

Martyniuk, T.; Puy, G.; Boulch, A.; Marlet, R.; and de Charette, R. 2025. LiDPM: Rethinking Point Diffusion for Lidar Scene Completion. *arXiv preprint arXiv:2504.17791*.

Milioto, A.; Vizzo, I.; Behley, J.; and Stachniss, C. 2019a. RangeNet++: Fast and accurate LiDAR semantic segmentation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4213–4220.

Milioto, A.; Vizzo, I.; Behley, J.; and Stachniss, C. 2019b. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 4213–4220. IEEE.

Mo, S.; Xie, E.; Chu, R.; Hong, L.; Niessner, M.; and Li, Z. 2023. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *Advances in neural information processing systems*, 36: 67960–67971.

Muhammad, K.; Hussain, T.; Ullah, H.; Ser, J. D.; Rezaei, M.; Kumar, N.; Hijji, M.; Bellavista, P.; and de Albuquerque, V. H. C. 2022. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Trans. Intell. Transport. Syst.*, 23(12): 22694–22715.

Nakashima, K.; and Kurazume, R. 2024. LiDAR Data Synthesis with Denoising Diffusion Probabilistic Models. In *IEEE International Conference on Robotics and Automation*, 14724–14731.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 8162–8171.

Nunes, L.; Marcuzzi, R.; Mersch, B.; Behley, J.; and Stachniss, C. 2024. Scaling Diffusion Models to Real-World 3D LiDAR Scene Completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14770–14780.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ran, H.; Guizilini, V.; and Wang, Y. 2024. Towards Realistic Scene Generation with LiDAR Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14738–14748.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; Vasudevan, V.; Han, W.; Ngiam, J.; Zhao, H.; Timofeev, A.; Ettinger, S.; Krivokon, M.; Gao, A.; Joshi, A.; Zhang, Y.; Shlens, J.; Chen, Z.; and Anguelov, D. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2446–2454.

Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; and Han, S. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, 685–702.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wilson, J.; Song, J.; Fu, Y.; Zhang, A.; Capodieci, A.; Jayakumar, P.; Barton, K.; and Ghaffari, M. 2022. MotionSC: Data Set and Network for Real-Time Semantic Mapping in Dynamic Environments. *IEEE Robotics and Automation Letters*, 7(3): 8439–8446.

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

Xie, H.; Chen, Z.; Hong, F.; and Liu, Z. 2025. CityDreamer4D: Compositional Generative Model of Unbounded 4D Cities. *arXiv preprint arXiv:2501.08983*.

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

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Zhai, G.; Örnek, E. P.; Chen, D. Z.; Liao, R.; Di, Y.; Navab, N.; Tombari, F.; and Busam, B. 2024. Echoscene: Indoor scene generation via information echo over scene graph diffusion. In *European Conference on Computer Vision*, 167–184. Springer.

Zhai, G.; Örnek, E. P.; Wu, S.-C.; Di, Y.; Tombari, F.; Navab, N.; and Busam, B. 2023. CommonScenes: generating commonsense 3D indoor scenes with scene graph diffusion. In *Advances in Neural Information Processing Systems*, 30026–30038.

Zhao, A.; Zhang, S.; Yang, L.; Li, Z.; Wu, J.; Xu, H.; Wei, A.; GU, P. P.; and Sun, L. 2025. Diffusion Distillation With Direct Preference Optimization For Efficient 3D LiDAR Scene Completion. *arXiv preprint arXiv:2504.11447*.

Zheng, Z.; Lu, F.; Xue, W.; Chen, G.; and Jiang, C. 2024. LiDAR4D: Dynamic Neural Fields for Novel Space-time View LiDAR Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5145–5154.

Zyrianov, V.; Zhu, X.; and Wang, S. 2022. Learning to generate realistic lidar point clouds. In *European Conference on Computer Vision*, 17–35. Springer.