Text2LiDAR: Text-guided LiDAR Point Cloud Generation via Equirectangular Transformer

Yang Wu[1], Kaihua Zhang[4], Jianjun Qian[1], Jin Xie[2,3†], and Jian Yang[1]

[1] PCA Lab, Nanjing University of Science and Technology, Nanjing, China
[2] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[3] School of Intelligence Science and Technology, Nanjing University, Suzhou, China
[4] B-DAT and CICAEET, Nanjing University of Information Science and Technology, Nanjing, China
{wuyang98,csjqian,csjxie,csjyang}@njust.edu.cn zhkhua@gmail.com

Abstract. The complex traffic environment and various weather conditions make the collection of LiDAR data expensive and challenging. Achieving high-quality and controllable LiDAR data generation is urgently needed, controlling with text is a common practice, but there is little research in this field. To this end, we propose Text2LiDAR, the first efficient, diverse, and text-controllable LiDAR data generation model. Specifically, we design an equirectangular transformer architecture, utilizing the designed equirectangular attention to capture LiDAR features in a manner with data characteristics. Then, we design a control-signal embedding injector to efficiently integrate control signals through the global-to-focused attention mechanism. Additionally, we devise a frequency modulator to assist the model in recovering high-frequency details, ensuring the clarity of the generated point cloud. To foster development in the field and optimize text-controlled generation performance, we construct nuLiDARtext which offers diverse text descriptors for 34,149 LiDAR point clouds from 850 scenes. Experiments on uncontrolled and text-controlled generation in various forms on KITTI-360 and nuScenes datasets demonstrate the superiority of our approach. The project can be found at https://github.com/wuyang98/Text2LiDAR

Keywords: LiDAR data generation · self-driving · diffusion models

1 Introduction

LiDAR provides accurate 3D geometry and distance information about the surroundings, enabling robots to understand the 3D environment. This capability makes LiDAR one of the most favored sensors in various autonomous systems, such as autonomous driving [12, 20, 26], unmanned surveying [13, 47], indoor exploration [79, 84, 88], to name a few. However, obtaining LiDAR data is not that straightforward. First, the price of LiDAR and its associated equipment is quite

† Corresponding author.

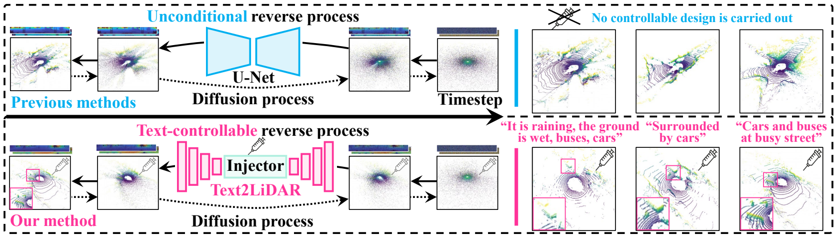
Unconditional reverse process
U-Net
Diffusion process
Previous methods
Timestep
No controllable design is carried out
Text-controllable reverse process
Injector
Text2LiDAR
Our method
Diffusion process
"It is raining, the ground is wet, buses, cars"
"Surrounded by cars"
"Cars and buses at busy street"

Figure 1: Schematic comparison of our Text2LiDAR and the existing diffusion-based generation framework [50, 89] without text guidance.

high [2]. Second, data collection in challenging situations presents safety and ethical concerns [15, 33, 52]. Therefore, high-quality generation of LiDAR point cloud [4, 14, 42, 48–50, 89] is becoming a frontier research area.

Existing works make a great effort in uncontrolled LiDAR point cloud generation that only uses single-modal LiDAR data. CARLA [14] starts from the physical meaning of LiDAR and simulates the imaging process. Due to significant disparities between physical models, CARLA cannot achieve satisfying performance. Afterward, Lidarsim [42] combines a physics-based and a learning-based simulation that can achieve better generations. However, it requires scanning real scenes in advance, which is time-consuming and labor-intensive. To address this issue, some asset-free methods [4, 48, 49] utilize the pure learning-based approaches, but they cannot fit more nonlinear distributions, limiting the diversity of the generation. Subsequent methods [50, 89] achieve improved results by employing U-Net [57, 58] as the denoising network in diffusion-based approaches. As illustrated in the upper part of Figure 1, the diffusion process can stimulate more complicated data distributions, thereby achieving more satisfying generations.

With the introduction of the CLIP [55] and the diffusion model [23, 62], the text-controlled generation tasks such as Text2image [29, 69, 76] and Text2video [67, 72] are developing rapidly. However, no works explore the text-guided paradigm in the field of LiDAR data generation. This primarily faces two challenges: **(1) No controllable generation architecture specially designed for equirectangular images and text.** Current leading methods [50, 89] all employ convolutional denoising architecture, represented by U-Net [57, 58]. The convolution architecture has two main limitations: first, ill-suiting for equirectangular images which have a circular structure, and convolutions disrupt the continuous relationship between pixels. Second, poor scalability, makes it inefficient and inconvenient to adapt to control signals from different modalities [24]. Besides, the existing methods [48, 50, 89] also overlook the correspondence between high-frequency information in equirectangular images and the structure of point cloud objects. These motivate us to explore how to build a unified controllable generation architecture that is compatible with the multi-modal signals of equirectangular images and text. **(2) No reliable text-LiDAR pairs for contrastive learning.** High-quality paired text-LiDAR data not only needs to describe the main objects in the LiDAR point cloud but also needs to include numerous diverse scenarios about weather, lighting, vehicle poses, and environmental structure to form a comprehensive description. Unfortunately, current mainstream datasets [3,5,18,37] cannot provide high-quality paired data. How to reasonably construct high-quality text-LiDAR data pair to adapt to the rapidly evolving field is also an important issue we aim to address.

In this paper, to address the challenge **(1)**, we present Text2LiDAR, a high-quality, text-controllable equirectangular transformer for LiDAR point cloud generation. As a usual practice [48, 50, 89], we convert each LiDAR scan into an equirectangular image and design targeted strategies based on its characteristics. We first design the equirectangular attention (EA) and reverse-EA (REA) for feature extraction and upsampling. They can capture long-range relationships between arbitrary two points, adapting to the circular structure of the equirectangular image, and addressing the disruption caused by convolution. Specifically, the EA introduces Fourier features to preserve 3D positional information while increasing differences between adjacent tokens for better model learning. Besides, the EA implements different-scale mutually overlapping unfolding operations to extract both global and local features, addressing the drastic deformations of objects caused by the elongated nature of equirectangular images. Then, to efficiently perform control signal fusion, we design a control-signal embedding injector (CEI) through a global-to-focused attention mechanism, endowing the model with text-controllable capabilities. At last, we design a frequency modulator (FM) to address the smooth characteristics of equirectangular images and overcome the smoothing tendency of MLP structures, ensuring the details of the generation. To overcome the challenge **(2)**, we construct a total of 34,149 pairs of high-quality text-LiDAR data across 850 scenes from the nuScenes [5], dubbed as nuLiDARtext. Based on the textual descriptions in nuScenes, we correct numerous abbreviations, spelling errors, and logical mistakes, and specifically adapt the text for LiDAR data. nuLiDARtext enhances the reliability of the text generation results and contributes to advancements in the field. The main technical contributions are summarized as follows:

- We propose the first effective text-controllable LiDAR point cloud generation framework, Text2LiDAR, which fully considers and adapts to the physical characteristics of the equirectangular image.
- We propose two novel module designs including a CEI and an FM. The CEI can progressively and robustly integrate control signals with dominant features through the global-to-focused attention mechanism, while the FM addresses the smoothing characteristics of equirectangular images and assists in model training, enhancing the quality of generation.
- To advance the field of LiDAR point cloud generation, nuLiDARtext is constructed, comprising 34,149 pairs of text-LiDAR data across 850 scenes.

2 Related Work

Point Cloud Generation. LiDAR point cloud generation is a subset of point cloud generation, emphasizing the generation of point cloud in autonomous driving scenarios [33, 68]. There is a strong correlation between the two tasks, due to not having to consider the surrounding scenes, the point cloud generation has earlier and more extensive related research. Earlier methods [14, 42] often rely on physical models, which makes them constrained by LiDAR equipment and capable only of achieving coarse generation. Sever works [1, 4, 25, 32, 65, 86] utilize the representative generative models such as generative adversarial networks (GANs) [19] and variational autoencoders (VAEs) [31] to solve point cloud generation. In addition, diverse generation methods have been proposed, achieving some effectiveness. Wu *et al.* [71] design a dual-generator framework that progressively extends the traditional GAN. SnowflakeNet [75] models the generation process as a snowflake-like growth, each point is generated from the original point after snowflake point deconvolution. Pointflow [82] introduces a two-level distribution structure, different levels represent different types of knowledge, allowing the model to sample point clouds of different sizes. Lou *et al.* [41] utilize diffusion techniques for generating point clouds, enabling high-quality point clouds with diverse scales. Gecco [64] improves the geometric consistency of point clouds by projecting sparse image features into the point cloud and using them as conditioning during the denoising process. By utilizing distillation techniques [22, 35, 36], Wu *et al.* [73] shorten the straight path to a single step, shorting the generation time of the standard diffusion model.

The outdoor LiDAR point cloud is very irregular and sparse. Due to the correlation between depth and LiDAR [45,80,81], in the tasks related to autonomous driving, LiDAR point cloud is often converted into equirectangular images to overcome the unstructuredness and sparsity of LiDAR point cloud [6, 43, 45]. Nakashima *et al.* [48, 49] decompose the noised equirectangular images into denoised forms and their corresponding dropout probabilities, significantly improving the performance. LiDARGen [89] designs a masking strategy to simulate the ray-drop in LiDAR, achieving diverse size generation results, and verifying the feasibility of using diffusion models for LiDAR data generation. The current leading work, R2DM [50], designs a more mature diffusion framework and achieves significant performance improvement. Despite numerous advancements, the realism of generated LiDAR point clouds and their diversity remain relatively low, and the absence of an efficient architecture with strong feature fusion capabilities for text control is still an unresolved issue.

Text in Vision. Thanks to the massive paired image-text data and the clever and concise model design, CLIP [55] can provide the semantically rich joint text-image representation, demonstrates strong capabilities across numerous visual tasks, such as low-light image enhancement [78], open-vocabulary object detection [34], image style transfer [77]. Thanks to the abundance of large-scale paired text-image datasets [10, 21, 38, 61, 70] and the cross-modal representation capability of CLIP, the currently booming field of text-controlled image generation [11, 17, 29, 56, 58, 69, 76, 87] effectively leverage text embeddings as control signals to guide the entire image generation process and achieve amazing results. All of these demonstrate that text and image features can be effectively fused. Thanks to the training objectives involving text-image pairs, the text embeddings generated by CLIP possess richer semantic information, making them particularly suitable as control conditions for generation tasks [58].

In the 3D vision domain, there is a scarcity of paired datasets, and compared to image generation, the development of 3D generation is relatively slow. Some methods [8,9,27,53,74] leverage pre-trained models and utilize Nerf [44] or Gaussian splatting [28] to achieve 3D generation. Nevertheless, in complex, realistic, and diverse 3D environments, these approaches are not universally applicable to all 3D generation tasks. Constructing text-paired datasets tailored to specific 3D generation tasks and conducting research on this basis is therefore particularly urgent. Some efforts are made in this direction. Chen *et al.* [7] construct a text-shape dataset and achieve text-controlled shape generation based on it. Liu *et al.* [39] further improve this work by decoupling the shape and color predictions for learning features. However, there are currently no relevant text-LiDAR datasets and generation frameworks. This paper aims to address these issues.

3 Method

3.1 Preliminary

This section introduces the formulation of the denoising diffusion probabilistic model (DDPM) and the loss function. As shown in Figure 1, the DDPM employs a forward diffusion process to gradually destroy the data sample $\mathbf{x}$ by adding noise as evolving the timestep $t \in [0, 1]$ until it becomes pure Gaussian noise. It also contains a backward reverse process, which aims at predicting the noise in each timestep and converting the pure Gaussian noise back into the data $\mathbf{x}$. To be more specific, at the timestep t, we can obtain the noised sample $\mathbf{x}_t$ through $q(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\alpha_t\mathbf{x}, \sigma_t^2\mathbf{I})$, where $\mathbf{x}_t$ can be re-parameterized as: $\mathbf{x}_t = \alpha_t\mathbf{x} + \sigma_t\boldsymbol{\epsilon}_t$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\epsilon}_t$ is the noise that vary with timestep t. α_t and σ_t are hyperparameters that depend on timestep t following the α-cosine schedule [50], we set $\alpha_t = \texttt{cos}(\pi t/2)$, $\sigma_t = \texttt{sin}(\pi t/2)$. Under the assumption $\alpha_t^2 + \sigma_t^2 = 1$, the process of obtaining the intermediate noised sample x_s can be described as $q(\mathbf{x}_t|\mathbf{x}_s) = \mathcal{N}(\alpha_{t|s}\mathbf{x}_s, \sigma_{t|s}^2\mathbf{I})$, where $0 \leq s < t \leq 1$, $\alpha_{t|s} = \alpha_t/\alpha_s$ and $\sigma_{t|s}^2 = \sigma_t - \alpha_{t|s}\sigma_s$. The corresponding reverse process can be described as:

$$p(\mathbf{x}_s|\mathbf{x}_t) = q(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}). \tag{1}$$

After obtaining the noised sample $\mathbf{x}_t$, we need to design a denoiser $\texttt{Text2LiDAR}_{\boldsymbol{\varphi}}$ to predict the noise $\hat{\boldsymbol{\epsilon}}_t = \texttt{Text2LiDAR}_{\boldsymbol{\varphi}}(\mathbf{x}_t, t)$ at each timestep t. Then, the denoised results can be obtained through Equation 1. Completing the entire denoising process for each timestep t, we can yield the final generated result. We use the mean squared error (MSE) loss function for the training process:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}[||\boldsymbol{\epsilon}_t - \texttt{Text2LiDAR}_{\boldsymbol{\varphi}}(\mathbf{x}_t, t)||_2^2], \tag{2}$$

where $\boldsymbol{\varphi}$ means the learnable parameters. As is customary [50], our denoiser is also conditioned on t. After training, we can obtain the final generated results by recursively evaluating $p(\mathbf{x}_s|\mathbf{x}_t)$ through the process for $t = 1 \rightarrow 0$.

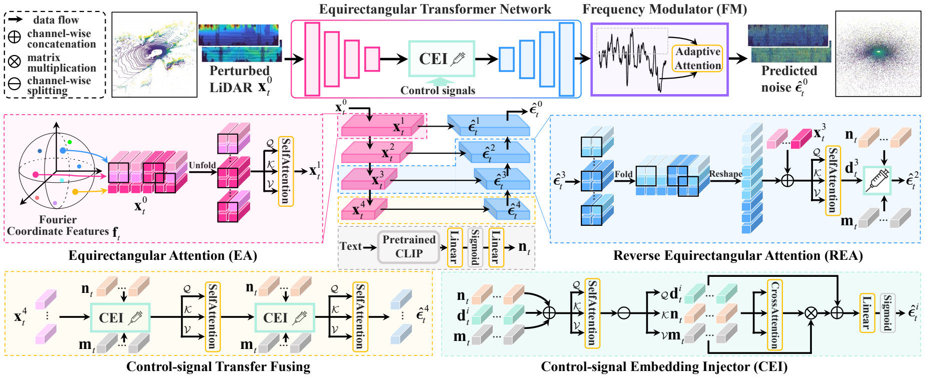

Figure 2: The architecture of the designed Text2LiDAR, where the designed equirectangular transformer is composed of stacked EA (encoding stage) and REA (decoding stage). The feature sequence will start interacting with the control signal at the 4th layer and be fed into a 4-layer decoder composed of REA. During decoding, the feature sequence continuously fuses the control signal through CEI. Finally, after frequency modulation, we can get the predicted noise.

3.2 Text2LiDAR Denoising Network

Figure 2 illustrates the architecture of our Text2LiDAR denoising network. At each timestep, the Text2LiDAR takes a noised equirectangular image $\mathbf{x}_t^0 \in \mathbb{R}^{H \times W \times 2}$ as input and outputs the predicted noise $\hat{\boldsymbol{\epsilon}}_t^0$ with the same size. The entire process is mainly composed of three parts, an equirectangular transformer network, a control-signal embedding injector, and a frequency modulator.

Equirectangular Transformer Network. Unlike regular images, equirectangular images possess unique physical properties. To this end, we design the equirectangular attention (EA) and highly adapt equirectangular images in three aspects. First, equirectangular images have a circular structure, previous methods employing circular convolution [60] still have limitations in expanding the convolution boundaries. In contrast, we utilize self-attention to break through these boundaries, enabling the capture of long-range relationships between arbitrary points. Second, pixels in the equirectangular image correspond to 3D positions. Previous methods only treated angular coordinates as additional input, resulting in minimal differences for adjacent regions, which is unfavorable for learning with MLP structures. Therefore, we utilize Fourier features [63] and extend the elevation and azimuth angles to frequency components of powers of two. This preserves the 3D priors while magnifying differences between neighboring positions, facilitating better model learning. Third, due to the elongated nature of the equirectangular image, the targets within it can undergo particularly exaggerated scale variations based on their distances, which are overlooked by previous methods. To address this, we use the mutually overlapping unfolding to cut the input sequence into different scales at different stages for local modeling. On these bases, our model can incorporate physical meanings to conduct

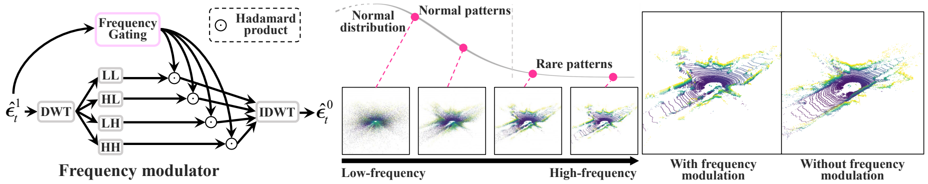

Figure 3: The architecture of the frequency modulator.

feature extraction comprehensively. The process can be written as:

$$\mathbf{x}_t^{i+1} = \texttt{EA}(\mathbf{x}_t^i) = \texttt{SelfAttention}(\texttt{Unfold}(\mathbf{x}_t^i \oplus \mathbf{f}_t)), \tag{3}$$

through the same steps, we can progressively obtain multi-level feature embeddings $\mathbf{x}_t^i \in \mathbb{R}^{H/2^i \times W/2^i \times C}, i = [1, 2, 3, 4]$.

Before decoding, as shown in the bottom-left corner of Figure 2, we process $\mathbf{x}_t^4$ using the control-signal transfer fusing for an initial fusion of control signals. In the decoding part of Text2LiDAR, we design reverse equirectangular attention (REA) for upsampling, allowing the continued capture of global-to-local relationships. For better recovering object details, we introduce features from the encoding stage. Simultaneously, to enhance the guidance of the embedding on the model, we use the designed control-signal embedding injector (CEI) at each upsampling stage to provide control. Through four stages of upsampling, we can upsample the token sequences to high resolution, matching the input size. This process can be written as:

$$\hat{\boldsymbol{\epsilon}}_t^i = \texttt{REA}(\mathbf{x}_t^{i+1}, \hat{\boldsymbol{\epsilon}}_t^{i+1}, \mathbf{n}_t, \mathbf{m}_t) = \texttt{CEI}(\texttt{SelfAttention}(\mathbf{x}_t^{i+1} \oplus \texttt{Fold}(\hat{\boldsymbol{\epsilon}}_t^{i+1})), \mathbf{n}_t, \mathbf{m}_t), \tag{4}$$

where $\mathbf{n}_t$ is the text embedding, $\mathbf{m}_t$ is the timestep embedding.

Control-signal Embedding Injector. One major reason previous methods can't achieve text-controlled generation is the inability to unify multi-modal features into tokens and integrate them. The bottom-right of Figure 2 illustrates the details of the CEI. Specifically, the dominant feature sequence $\mathbf{d}_t^i$ obtained from the preceding steps, text embedding $\mathbf{n}_t$ and timestep embedding $\mathbf{m}_t$ are first concatenated for self-attention operation, and the obtained feature sequence is then split in the order of concatenation to form sever new feature sequences $\mathcal{Q}\mathbf{d}_t^i, \mathcal{K}\mathbf{n}_t, \mathcal{V}\mathbf{m}_t$. These new sequences are for cross-attention operation [66,85]. To preserve the guiding effect of the timestep $\mathcal{V}\mathbf{m}_t$, we performed additional matrix multiplication for it. To consolidate the dominant denoising feature sequence $\mathbf{d}_t^i$, we add it to the feature sequence after multiplication, obtaining the final sequence $\hat{\boldsymbol{\epsilon}}_t^i$. This process can be formulated as:

$$\hat{\boldsymbol{\epsilon}}_t^i = \texttt{CrossAttention}(\texttt{Split}(\texttt{SelfAttention}(\mathbf{n}_t \oplus \mathbf{d}_t^i \oplus \mathbf{m}_t))) \otimes \mathcal{V}\mathbf{m}_t \oplus \mathcal{Q}\mathbf{d}_t^i, \tag{5}$$

This is a global-to-focused mechanism that, while integrating control signals, maintains the ability to generate key features from LiDAR data. It's worth mentioning that our designed injector can easily remove text embeddings. Only the composition of $\mathcal{QKV}$ needs to be adjusted.

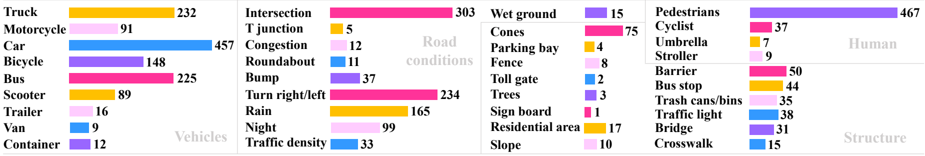

Figure 4: The number of occurrences of text in 850 scenes.

Frequency Modulator. As shown in Figure 3, in the diffusion model, low-frequency information is restored first, followed by the gradual restoration of high-frequency information [83]. The high-frequency information in equirectangular images affects generation details, but it is prone to being smoothed out by MLP operations To this end, we design a frequency modulator (FM) to allow the model to adaptively focus on high-frequency information. The left side of Figure 3 illustrates its structure, mainly composed of a discrete wavelet transform (DWT), a frequency gating function (FG) composed of convolutions, and an inverse discrete wavelet transform (IDWT) [16, 83]. The goal is to decompose the input $\hat{\boldsymbol{\epsilon}}_t^1$ into multi-angle high-frequency wavelet bands for modulation, guiding the model to adapt adaptively to different frequencies, alleviating the transition smoothness of the equirectangular image. The process can be written as:

$$\hat{\boldsymbol{\epsilon}}_t^0 = \mathtt{FM}(\hat{\boldsymbol{\epsilon}}_t^1) = \mathtt{IDWT}(\mathtt{DWT}(\hat{\boldsymbol{\epsilon}}_t^1) \odot \mathtt{FG}(\hat{\boldsymbol{\epsilon}}_t^1)). \tag{6}$$

We can see from the right side of Figure 3, that with the designed FM, the details of the LiDAR point clouds are clearer. In the end, we obtain the predicted denoised results and perform end-to-end model training using Equation 2.

3.3 Construction of NuLiDARtext

Enabling text control can significantly enhance the practicality of LiDAR point clouds generation, meeting personalized customization needs for weather and road conditions. We introduce nuLiDARtext to promote the advancement of the field. To save resources and costs, we construct text descriptions suitable for single-frame LiDAR point cloud generation on existing nuScenes [5]. The text descriptions in the nuScenes dataset are intended for describing scenes within a short period and are not paired specifically for LiDAR data. When attempting to pair text data with LiDAR data, we find issues such as misspellings, semantic redundancies, continuous state descriptions, and interference words. Therefore, we manually adjust the descriptions of 34,149 frames across 850 scenes provided by nuScenes, including operations such as addition, deletion, modification, and standardization. For example, we correct "ped" to "pedestrians" since this abbreviation is not conducive to obtaining effective text embedding. We also adjust instances where both "turn left" and "turn right" are simultaneously mentioned to more realistically represent "turn left". Additionally, we remove "hidden ped", which represents information not perceivable by LiDAR. We modify "waiting at the intersection" to "at the intersection", as "waiting" is a continuous state description that could dilute the effective information for a single-frame generation. We make corrections and modifications to almost every text description. The proposed nuLiDARtext can better provide paired text-LiDAR data, promoting advancements in the field. Figure 4 illustrates the main components of the current textual prompts, and the presentation of the dataset can be found in the supplementary materials.

4 Experiments

4.1 Implementation Details

Our model is built using PyTorch [51] and trained on 4 NVIDIA RTX 3090 GPUs, requiring a total of 300,000 training steps. Our model is optimized using the Adam algorithm [30] with exponential decay rates of 0.9 and 0.99, and the learning rate is set to $1e-4$.

4.2 Datasets and Evaluation Metrics

We conduct experiments on two challenging datasets, KITTI-360 and nuScenes. KITTI-360 [37] provides 360-degree, 64-beam LiDAR point clouds, allowing intelligent devices to comprehensively understand the 3D structure of the surroundings. As usual practice [50, 89], we split KITTI-360 into two parts, with one portion (50348 frames) used for training and validation, and the other portion (26367 frames) reserved for testing. When projecting onto equirectangular images, we set the dimensions of the image to $64 \times 1024 \times 2$. NuScenes [5] provides continuous data collection for 20 seconds, and we primarily utilize its "samples" data, which consists of 34,149 frames of 360-degree, 32-beam LiDAR point clouds. The projected equirectangular images are set to $32 \times 1024 \times 2$.

Consistent with LiDARGen [89] and R2DM [50], for unconditional generation, we compute the distributional dissimilarity between 10,000 generated samples and real samples, conducting evaluations in three data formats: image, point clouds, and bird's eye view (BEV). In equirectangular image form, we utilize the pre-trained RangeNet [46] to calculate the Frechet range distance (FRD) [89]. In point cloud form, we use the pre-trained PointNet [54] to obtain the Frechet point cloud distance (FPD) metric, which calculates similarly to FRD, to evaluate the difference between generated samples and real samples at the point cloud level. In BEV form, we use Jensen–Shannon divergence (JSD) and minimum matching distance (MMD) to measure the distance between the marginal distributions of BEV occupancy grids.

4.3 LiDAR Uncontrolled Generation

In this section, we conducted quantitative and qualitative analyses of the results of uncontrolled generation. We compare our approach with three GAN [4, 59] or VAE methods [4] and two diffusion methods [50,89]. Among them, the evaluation

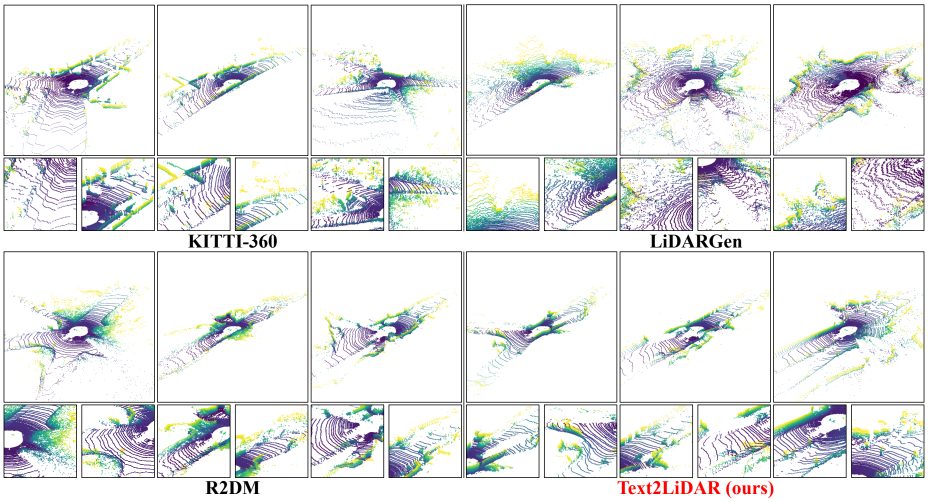

Figure 5: Comparison with LiDARGen and R2DM on uncontrolled generation.

Table 1: Comparison of four metrics with state-of-the-art methods on KITTI-360.

Method	Base	T	Point cloud	Image	BEV occupancy grid	
			FPD↓	FRD↓	MMD$\times 10^{-4}\downarrow$	JSD$\times 10^{-2}\downarrow$
LiDARGAN* [4]	GAN	None	-	3003.8	30.60	-
LiDARVAE* [4]	VAE	None	-	2261.5	10.00	16.10
ProjectedGAN* [59]	GAN	None	-	2117.2	3.47	8.50
LiDARGen [89]	NCSNv2	1160	90.29	579.39	7.39	7.38
R2DM [50]	DDPM	256	6.24	**149.66**	1.91	3.05
Text2LiDAR (ours)	DDPM	256	**4.81**	164.16	**0.49**	**2.01**

metrics for methods marked with "*" are provided by LiDARGen [89], while other methods are trained with uniform settings from their source code and evaluated accordingly, and we standardize R2DM to be trained in single precision to keep the same with LiDARGen and Text2LiDAR.

Table 1 presents the experimental results. It can be observed that our approach achieves the best performance in terms of FPD, MMD, and JSD. This indicates that the point clouds generated by our model exhibit the most realistic distribution in both the 3D and 2D planes. Our method also performs well in the FRD metric, with a slight difference compared to R2DM. However, this does not affect our ability to generate high-quality LiDAR point clouds.

Figure 5 displays the uncontrolled generation results of our method and the compared methods. We can see that LiDARGen exhibits the basic characteristics of LiDAR point clouds but suffers from noticeable blurriness and more noise. Compared to R2DM, our method better captures the realistic state of distant

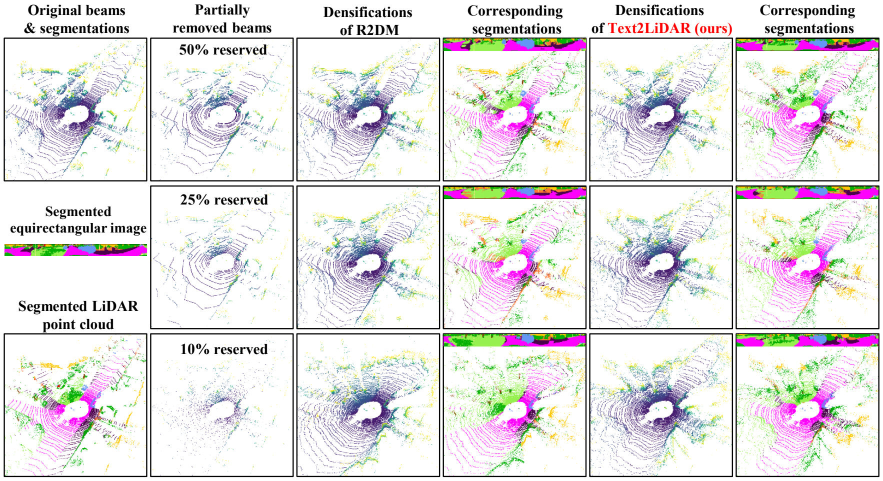
Original beams & segmentations
Partially removed beams
Densifications of R2DM
Corresponding segmentations
Densifications of Text2LiDAR (ours)
Corresponding segmentations
50% reserved
25% reserved
10% reserved
Segmented equirectangular image
Segmented LiDAR point cloud

Figure 6: Comparison with R2DM on densification.

Table 2: Comparison of densification performance with R2DM on KITTI-360.

Method	Depth		Intensity		Semantic
	MAE↓	RMSE↓	MAE↓	RMSE↓	IoU% ↑
R2DM	0.039	0.114	0.090	0.146	40.16
Text2LiDAR(ours)	0.035	0.102	0.073	0.134	41.37

beams, owing to the effective capturing of the LiDAR features from equirectangular images by our designed EA. Thanks to the designed frequency modulator, the generation of target outlines and boundaries is also clearer in our method. We present more generated results in supplementary materials

4.4 LiDAR Densification

Compared to directly generating LiDAR point clouds in uncontrolled manners, densifying existing sparse LiDAR point clouds (32,16 beams) is also an effective method for data generation. It can significantly reduce data collection costs, and can effectively enhance the accuracy and safety of unmanned systems [37]. To maintain consistency with the baseline, we use the framework designed based on R2DM [50] on top of Repaint [40] for densification. In addition to observing the completeness and realism of the LiDAR point clouds, we also need to verify the rationality. For effectively evaluating, we employ RangeNet [46] for semantic segmentation on both the initial and the densification, aiming to observe whether the consistency in semantic information has been well preserved.

As shown in Figure 6, in the first column, we present the original LiDAR point clouds and their segmentation results. In the second column, we showcase the states of the original LiDAR point cloud with a reduction of 50% and 75%

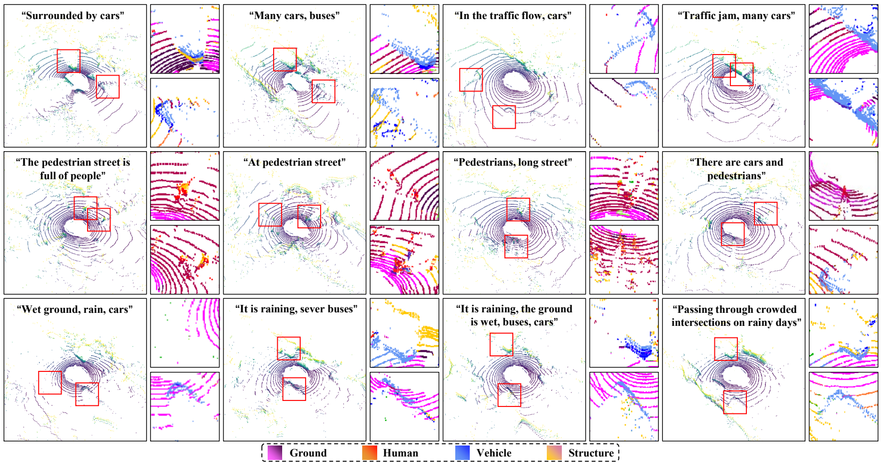
"Surrounded by cars"
"Many cars, buses"
"In the traffic flow, cars"
"Traffic jam, many cars"
"The pedestrian street is full of people"
"At pedestrian street"
"Pedestrians, long street"
"There are cars and pedestrians"
"Wet ground, rain, cars"
"It is raining, sever buses"
"It is raining, the ground is wet, buses, cars"
"Passing through crowded intersections on rainy days"
Ground
Human
Vehicle
Structure

Figure 7: Results of text-controlled LiDAR point clouds generation.

of beams, as well as the state with 90% of points randomly removed. It is not difficult to see that the latter two are quite challenging.

In the third and fourth columns of Figure 6, we show the results of densification and semantic segmentation of R2DM, and in the next two columns, we present the results of our Text2LiDAR. It is evident that the point clouds densified by our method align more closely with the characteristics of the real, and the generation of semantic information is more complete and accurate. Taking the most challenging case in the third row as an example, R2DM exhibits large-scale misjudgments for the road below and on the left, as well as misjudgments for the vegetation on the left. In contrast, our method better preserves the road surface and vegetation, aligning more closely with the characteristics of initial point clouds.

For quantitative comparison, we test 100 samples on KITTI-360. We utilize mean absolute error (MAE) and root mean squared error (RMSE) as metrics to assess the quality of isotropic equirectangular image densification. Additionally, intersection over union (IoU) is employed to measure the ability to preserve semantic information. Table 2 shows the performance comparison between our method and R2DM under the condition of randomly removing 90% of points. We can see that our method outperforms R2DM across all metrics. This is attributed to the designed EA and REA, which perform feature extraction and upsampling based on the physical characteristics of the equirectangular image

4.5 Text-controlled Generation

Previous methods are all limited to unconditional generation and densification, only our approach enables text-controlled generation, which can further expand application scenarios. Thanks to our designed CEI, our model can effectively integrate textual semantics into the generation process. Figure 7 illustrates the

Table 3: Results of ablation study on key designs.

Versions	Key designs					Point cloud	Image	BEV occupancy grid	
	EA	REA	CEI	FM	D-layers	FPD↓	FRD↓	MMD$\times 10^{-4}$ ↓	JSD$\times 10^{-2}$ ↓
Text2LiDAR$_A$	-	-	-	-	3	25.03	692.15	1.26	3.82
Text2LiDAR$_B$	✓	-	✓	✓	4	8.02	305.27	0.81	2.90
Text2LiDAR$_C$	-	✓	✓	✓	4	7.41	220.11	0.87	3.09
Text2LiDAR$_D$	-	-	✓	✓	4	10.52	416.03	0.89	3.30
Text2LiDAR$_E$	✓	✓	-	✓	4	6.99	187.82	0.83	2.79
Text2LiDAR$_F$	✓	✓	✓	-	4	5.95	165.33	0.52	2.45
Text2LiDAR$_G$	✓	✓	-	-	4	11.15	333.69	1.17	4.18
Text2LiDAR$_H$	✓	✓	✓	✓	3	19.36	521.32	0.97	3.59
Text2LiDAR	✓	✓	✓	✓	4	**4.81**	**164.16**	**0.49**	**2.01**

text-controlled 32-beam generations. To verify the semantic accuracy of the generated objects, we employ RangeNet [46] for semantic segmentation, and the zoomed-in results are displayed in the right column. Additional diverse text-controlled generation results are also showcased in the supplementary materials

Larger Objects Generation. It can be observed from the first row that the generated outlines of vehicles are very clear, and the model is capable of generating various forms of vehicles, including small cars, buses, and trucks. The distribution of LiDAR beams also aligns with the characteristics of real data.

Human Generation. Generating humans is particularly challenging, as the LiDAR points that successfully reflect off and are detected on a human body in an outdoor scene are quite sparse. It can be observed from the second row that our method is capable of generating high-quality representations of humans while maintaining the coherence of other parts of the point cloud.

Rainy Day Generation. The third row shows the more challenging and meaningful results of generating LiDAR point clouds on rainy days. Due to the wet surfaces of the ground and objects from rain, the beams illuminating them experience total reflection, making it challenging for LiDAR to detect them. This leads to sparser or even zero returning beams from targets that are farther away from the LiDAR. Particularly noteworthy is the third example in the third row, which illustrates a typical phenomenon where beams that illuminate the vehicle can return, while those illuminating the ground around the vehicle cannot.

4.6 Ablation Study and Analysis

We conducted experiments on KITTI-360, varying the effectiveness of the main designs in this paper, including EA, REA, CEI, FM, and the number of decoder layers (D-layers). From Table 3, we can see the experimental results.

Benefits of (Reverse) Equirectangular Attention. Table 3 indicates that when we do not use EA or REA and only use the original self-attention as a substitute, the model's performance deteriorates. The performance drop is more pronounced when REA is not used, as upsampling is closer to the model's endpoint, directly impacting high-resolution output. When both REA and EA are not used,

Table 4: Comparison with the parameters and generation speed of previous methods.

Methods	Network architecture	Control signal	Parameters(M)	Time(s)
LiDARGen [89]	RefineNet	None	29.7	88.54
R2DM [50]	Efficient U-Net	None	31.1	7.79
Text2LiDAR(ours)	Transformer	None	45.8	12.57
Text2LiDAR(ours)	Transformer	Text	46.1	4.57

the model's performance further declines, deteriorating by $5.71, 251.87, 0.4 \times 10^{-4}$, and 1.29×10^{-2}, in the respective four metrics. Meanwhile, we find that increasing the number of decoding layers is an effective way to enhance performance, further confirming the effectiveness of REA.

Benefits of Control-signal Embedding Injector. In terms of feature fusion, we compare channel-wise concatenation with the designed CEI. We can see from Text2LiDAR$_E$ that, in addition to fusing text embedding, the designed injector is also more effective in incorporating timestep embedding, ensuring the stability of the reverse process.

Benefits of Frequency Modulator. From Text2LiDAR$_F$, we can find that, without FM, all four metrics show varying degrees of decline: $0.78, 1.17, 0.03 \times 10^{-4}$, and 0.44×10^{-2}. Furthermore, we observe that FM has a relatively minor impact on FRD. We attribute this to the small proportion of high-frequency information, and adjusting it could introduce some image distortion, which is crucial for the performance of point clouds.

Parameter Analysis. Table 4 shows the comparison of model parameters and generation speed between our method, R2DM [50], and LiDARGen [89]. Our model has two versions, distinguished based on the control signal. It can be observed that our method has a larger number of parameters, but this does not hinder the ability to perform fast generating. When generating 32-line LiDAR data with text control, our model achieves high speed, requiring less than five seconds. This indicates that our model is highly practical.

5 Conclusion

We have proposed Text2LiDAR, the first high-performance text-controllable LiDAR data generation framework. With the designed EA, Text2LiDAR can specifically capture LiDAR features from equirectangular images while fully considering the physical meaning of equirectangular images. Then, a CEI has been proposed to integrate control signals , ensuring high-quality text-controlled generation. We have further proposed an FM to overcome the issue of excessive smoothing in equirectangular images and assist the model training. Our Text2LiDAR has shown great performance in uncontrolled and text-controlled generation, and densification, with promising applications across various domains. At last, we have constructed the nuLiDARtext, providing the paired text-LiDAR data to facilitate advancements in this field.

Acknowledgements

We are very grateful for the reviewers' critical and constructive comments. This work was supported in part by the NSFC under Grant No. 62276141, 62176124, 62276144, 62361166670.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: International conference on machine learning. pp. 40–49. PMLR (2018)
2. Bakhshi, R., Sandborn, P.: Maximizing the returns of lidar systems in wind farms for yaw error correction applications. Wind Energy **23**(6), 1408–1421 (2020)
3. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In: Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV) (2019)
4. Caccia, L., Van Hoof, H., Courville, A., Pineau, J.: Deep generative modeling of lidar data. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5034–5040. IEEE (2019)
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
6. Chai, Y., Sun, P., Ngiam, J., Wang, W., Caine, B., Vasudevan, V., Zhang, X., Anguelov, D.: To the point: Efficient 3d object detection in the range image with graph convolution kernels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2021)
7. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 100–116. Springer (2019)
8. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
9. Chen, Z., Wang, F., Liu, H.: Text-to-3d using gaussian splatting. arXiv preprint arXiv:2309.16585 (2023)
10. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3043–3054 (2023)
11. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: Vqgan-clip: Open domain image generation and editing with natural language guidance. In: European Conference on Computer Vision. pp. 88–105. Springer (2022)
12. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al.: A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 958–979 (2024)

13. Deliry, S.I., Avdan, U.: Accuracy of unmanned aerial systems photogrammetry and structure from motion in surveying and mapping: a review. Journal of the Indian Society of Remote Sensing **49**(8), 1997–2017 (2021)
14. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
15. Dreissig, M., Scheuble, D., Piewak, F., Boedecker, J.: Survey on lidar perception in adverse weather conditions. arXiv preprint arXiv:2304.06312 (2023)
16. Fu, M., Liu, H., Yu, Y., Chen, J., Wang, K.: Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–212 (2021)
17. Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7545–7556 (2023)
18. Geiger, A., Lenz, P., Urtasun, R.: 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)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
20. Gulino, C., Fu, J., Luo, W., Tucker, G., Bronstein, E., Lu, Y., Harb, J., Pan, X., Wang, Y., Chen, X., et al.: Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. Advances in Neural Information Processing Systems **36** (2024)
21. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019)
22. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
25. Hui, L., Xu, R., Xie, J., Qian, J., Yang, J.: Progressive point cloud deconvolution generation network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 397–413. Springer (2020)
26. Janai, J., Güney, F., Behl, A., Geiger, A., et al.: Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision **12**(1–3), 1–308 (2020)
27. Kasten, Y., Rahamim, O., Chechik, G.: Point cloud completion with pretrained text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
29. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711 (2023)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

31. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
32. Klokov, R., Boyer, E., Verbeek, J.: Discrete point flow networks for efficient point cloud generation. In: European Conference on Computer Vision. pp. 694–710. Springer (2020)
33. Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3d: Towards robust and reliable 3d perception against corruptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19994–20006 (2023)
34. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639 (2022)
35. Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., Yang, J.: Promptkd: Unsupervised prompt distillation for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26617–26626 (2024)
36. Li, Z., Li, X., Yang, L., Zhao, B., Song, R., Luo, L., Li, J., Yang, J.: Curriculum temperature for knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1504–1512 (2023)
37. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3292–3310 (2022)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
39. Liu, Z., Wang, Y., Qi, X., Fu, C.W.: Towards implicit text-guided 3d shape generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17896–17906 (2022)
40. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
41. Luo, S., Hu, W.: 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)
42. Manivasagam, S., Wang, S., Wong, K., Zeng, W., Sazanovich, M., Tan, S., Yang, B., Ma, W.C., Urtasun, R.: Lidarsim: Realistic lidar simulation by leveraging the real world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11167–11176 (2020)
43. Meyer, G.P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., Wellington, C.K.: Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12677–12686 (2019)
44. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
45. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 4213–4220. IEEE (2019)

46. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 4213–4220. IEEE (2019)
47. Mohsan, S.A.H., Othman, N.Q.H., Li, Y., Alsharif, M.H., Khan, M.A.: Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends. Intelligent Service Robotics **16**(1), 109–137 (2023)
48. Nakashima, K., Iwashita, Y., Kurazume, R.: Generative range imaging for learning scene priors of 3d lidar data. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1256–1266 (2023)
49. Nakashima, K., Kurazume, R.: Learning to drop points for lidar scan synthesis. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 222–229. IEEE (2021)
50. Nakashima, K., Kurazume, R.: Lidar data synthesis with denoising diffusion probabilistic models. arXiv preprint arXiv:2309.09256 (2023)
51. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
52. Piroli, A., Dallabetta, V., Kopp, J., Walessa, M., Meissner, D., Dietmayer, K.: Energy-based detection of adverse weather effects in lidar data. IEEE Robotics and Automation Letters (2023)
53. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
54. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
56. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
57. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
58. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
59. Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected gans converge faster. Advances in Neural Information Processing Systems **34**, 17480–17492 (2021)
60. Schubert, S., Neubert, P., Pöschmann, J., Protzel, P.: Circular convolutional neural networks for panoramic images and laser data. In: 2019 IEEE intelligent vehicles symposium (IV). pp. 653–660. IEEE (2019)
61. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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)

62. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
63. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems **33**, 7537–7547 (2020)
64. Tyszkiewicz, M.J., Fua, P., Trulls, E.: Gecco: Geometrically-conditioned point diffusion models. arXiv preprint arXiv:2303.05916 (2023)
65. Valsesia, D., Fracastoro, G., Magli, E.: Learning localized generative models for 3d point clouds via graph convolution. In: International conference on learning representations (2018)
66. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
67. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
68. Wang, Y., Mao, Q., Zhu, H., Deng, J., Zhang, Y., Ji, J., Li, H., Zhang, Y.: Multi-modal 3d object detection in autonomous driving: a survey. International Journal of Computer Vision pp. 1–31 (2023)
69. Wang, Z., Liu, W., He, Q., Wu, X., Yi, Z.: Clip-gen: Language-free training of a text-to-image generator with clip. arXiv preprint arXiv:2203.00386 (2022)
70. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.14896 (2022)
71. Wen, C., Yu, B., Tao, D.: Learning progressive point embeddings for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10266–10275 (2021)
72. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
73. Wu, L., Wang, D., Gong, C., Liu, X., Xiong, Y., Ranjan, R., Krishnamoorthi, R., Chandra, V., Liu, Q.: Fast point cloud generation with straight flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9445–9454 (2023)
74. Wu, Z., Wang, Y., Feng, M., Xie, H., Mian, A.: Sketch and text guided diffusion model for colored point cloud generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8929–8939 (2023)
75. Xiang, P., Wen, X., Liu, Y.S., Cao, Y.P., Wan, P., Zheng, W., Han, Z.: Snowflake point deconvolution for point cloud completion and generation with skip-transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 6320–6338 (2022)
76. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
77. Xu, Z., Xing, S., Sangineto, E., Sebe, N.: Spectralclip: Preventing artifacts in text-guided style transfer from a spectral perspective. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5121–5130 (2024)
78. Xue, M., He, J., He, Y., Liu, Z., Wang, W., Zhou, M.: Low-light image enhancement via clip-fourier guided wavelet diffusion. arXiv preprint arXiv:2401.03788 (2024)

79. Yan, Z., Li, X., Wang, K., Zhang, Z., Li, J., Yang, J.: Multi-modal masked pretraining for monocular panoramic depth completion. In: European Conference on Computer Vision. pp. 378–395. Springer (2022)
80. Yan, Z., Lin, Y., Wang, K., Zheng, Y., Wang, Y., Zhang, Z., Li, J., Yang, J.: Triperspective view decomposition for geometry-aware depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4874–4884 (2024)
81. Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J., Yang, J.: Rignet: Repetitive image guided network for depth completion. In: European Conference on Computer Vision. pp. 214–230. Springer (2022)
82. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4541–4550 (2019)
83. Yang, X., Zhou, D., Feng, J., Wang, X.: Diffusion probabilistic model made slim. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22552–22562 (2023)
84. Yin, H., Lin, Z., Yeoh, J.K.: Semantic localization on bim-generated maps using a 3d lidar sensor. Automation in Construction **146**, 104641 (2023)
85. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 558–567 (2021)
86. Zamorski, M., Zięba, M., Klukowski, P., Nowak, R., Kurach, K., Stokowiec, W., Trzciński, T.: Adversarial autoencoders for compact representations of 3d point clouds. Computer Vision and Image Understanding **193**, 102921 (2020)
87. Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., Sun, T.: Towards language-free training for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17907–17917 (2022)
88. Zou, Q., Sun, Q., Chen, L., Nie, B., Li, Q.: A comparative analysis of lidar slam-based indoor navigation for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems **23**(7), 6907–6921 (2021)
89. Zyrianov, V., Zhu, X., Wang, S.: Learning to generate realistic lidar point clouds. In: European Conference on Computer Vision. pp. 17–35. Springer (2022)