A Self-Conditioned Representation Guided Diffusion Model for Realistic Text-to-LiDAR Scene Generation

Wentao Qu[1], Guofeng Mei[2], Yang Wu[1], YongShun Gong[3], Xiaoshui Huang[4*], Liang Xiao[1*]
NJUST[1], FBK[2], SDU[3], SJTU[4]
quwentao@njust.edu.cn, huangxiaoshui@163.com, xiaoliang@mail.njust.edu.cn

Abstract

*Text-to-LiDAR generation can customize 3D data with rich structures and diverse scenes for downstream tasks. However, the scarcity of Text-LiDAR pairs often causes insufficient training priors, generating overly smooth 3D scenes. Moreover, low-quality text descriptions may degrade generation quality and controllability. In this paper, we propose a **T**ext-**t**o-**L**iDAR **D**iffusion **M**odel for scene generation, named T2LDM, with a Self-Conditioned Representation Guidance (SCRG). Specifically, SCRG, by aligning to the real representations, provides the soft supervision with reconstruction details for the Denoising Network (DN) in training, while decoupled in inference. In this way, T2LDM can perceive rich geometric structures from data distribution, generating detailed objects in scenes. Meanwhile, we construct a content-composable Text-LiDAR benchmark, T2nuScenes, along with a controllability metric. Based on this, we analyze the effects of different text prompts for LiDAR generation quality and controllability, providing practical prompt paradigms and insights. Furthermore, a directional position prior is designed to mitigate street distortion, further improving scene fidelity. Additionally, by learning a conditional encoder via frozen DN, T2LDM can support multiple conditional tasks, including Sparse-to-Dense, Dense-to-Sparse, and Semantic-to-LiDAR generation. Extensive experiments in unconditional and conditional generation demonstrate that T2LDM outperforms existing methods, achieving state-of-the-art scene generation.*

1. Introduction

LiDAR perceives the surrounding environment and geometric structures, providing essential data support for 3D scene understanding tasks such as autonomous driving [43], virtual reality [55], and robotics [1]. However, collecting LiDAR scene data with diverse structures and under adverse weather conditions (e.g., rain) is costly [3, 36, 55], limiting the advance of data-driven 3D perception models. Therefore, this has motivated growing research interest in synthesizing realistic, diverse and controllable LiDAR scenes.

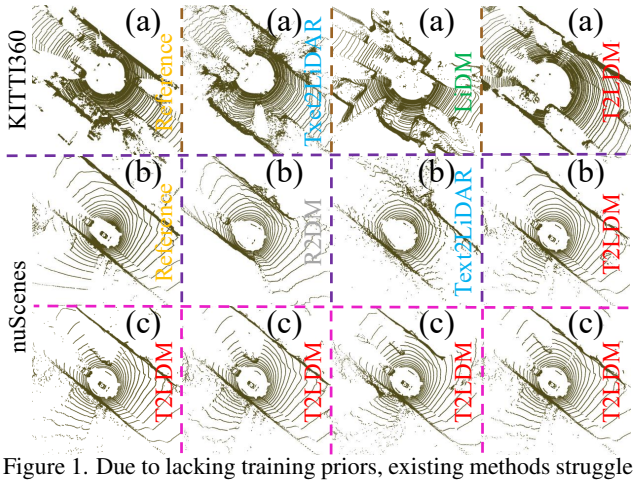

Figure 1. Due to lacking training priors, existing methods struggle to generate detailed scene objects. In contrast, T2LDM can generate realistic (a and b) and diverse (c) variants of the same scene.

Convenient natural language can provide semantic guidance for controllable scene generation. In recent years, numerous studies have achieved success in Text-to-Image generation tasks. Benefiting from rich Text-Image paired data [4, 50], some methods can be trained on even over 100 million samples [45, 47, 48]. This provides strong cross-modal alignment and semantic priors, enabling generative models to synthesize realistic and diverse visual content from natural language descriptions. Inspired by this, researchers attempt to introduce text conditions for customizing LiDAR scenes [55], improving 3D data diversity and scalability.

Unfortunately, unlike the easily collected Text-Image pairs (*e.g.*, the internet [4, 44]), LiDAR data acquisition is time-consuming and labor-intensive, making high-quality and diverse Text-LiDAR pairs extremely scarce (the Text-LiDAR pairs $<$ 35K in nuScenes [6]). This limitation hinders the sufficient training of generative models, thus often resulting in overly smooth and homogeneous generation results that lack distinct object structures and realistic details in LiDAR scenes (see Fig. 1). Moreover, the content and form of text prompts for describing scene structures are crucial to generation results [32, 38]. However, existing bench-

*Corresponding Author. https://github.com/QWTforGithub/T2LDM

marks [6] only provide unnatural text descriptions and lack controllability evaluation metrics, further limiting the generation results of Text-to-LiDAR generative models.

To address these problems, in this paper, we propose a Text-to-LiDAR Diffusion Model , named T2LDM. Inspired by injecting regularization into DDPMs through representation learning [23, 62], T2LDM employs a Self-Conditional Representation Guidance (SCRG) to learn geometric details from data distribution, improving object fidelity in generated LiDAR scenes. Specifically, SCRG leverages a Guidance Network (GN) to perceive multi-scale perturbed features from DN while aligning with the real representations. This allows GN to produce geometrically detailed features under multi-level perturbations and conditional guidance, providing multi-scale supervision signals for DN. In this way, T2LDM can effectively learn geometric structures from data distribution, generating detailed and realistic objects in scenes (see Fig. 1). Moreover, unlike requiring pretrained priors [23, 62], SCRG operates in an end-to-end paradigm, and ***GN participates in gradient backpropagation only during the early training stage while detached during inference,*** alleviating the computational cost.

Meanwhile, we construct a content-composable Text-LiDAR benchmark, T2nuScenes, with 3D box priors. This offers three advantages. First, this enables more accurate description of object locations than manual annotations (accuracy). Second, this can generalize to any 3D detection dataset (generality). Third, this enables controllability evaluation via detectors (evaluability). Moreover, ***Text-to-LiDAR generation presents more flexibility than conditioning on complex 3D boxes.*** This is also the first to explore 3D box priors for scene text description, encouraging more explorations of 3D text-guided scene generation. Based on this, we investigate the effect of text forms for LiDAR generation, providing prompt paradigms and insights.

Furthermore, we find that the spherical projection from LiDAR data to range map may cause ***directional confusion***, leading to distorted streets in generated scenes (see Fig. 2(c)). Therefore, we design a directional position encoding to provide T2LDM with true directional priors for rows and columns of range maps, further improving fidelity.

Additionally, T2LDM exhibits excellent results across various conditional generation tasks via ***non-latent ControlNet [64]***. Our key contributions can be summarized as:

- We propose a Text-to-LiDAR Diffusion Model, T2LDM, with a self-conditioned representation guidance.
- We construct a high-quality content-composable Text-LiDAR benchmark, T2nuScenes, exploring effective text prompt forms and providing insights.
- By leveraging a directional position prior, T2LDM alleviates road distortion, further improving scene fidelity.
- Unconditional and conditional results show that T2LDM can generate LiDAR scenes with detailed objects.

2. Related Works

LiDAR Scene Generation. LiDAR data can precisely describe the geometric structures and spatial relationships for a scene, providing an effective representation of the real-world environment. However, the high acquisition cost and labor-intensive annotation process make high-quality and diverse LiDAR data extremely scarce [5, 36, 55, 56, 71]. Some methods attempt to synthesize realistic scenes on LiDAR data using physics-based simulation [14, 34, 53, 59]. They typically model the physics of LiDAR signals based on optical scattering and laser propagation principles, simulating signal attenuation, backscattering, and measurement noise. Although effective, directly simulating realistic scenes on LiDAR data struggles to generate geometrically diverse and structurally rich scenes, as this requires high-quality LiDAR data as the shape foundation. Benefiting from the strong data-driven capability of deep learning [7–12, 19, 20, 24, 25, 28, 29, 51, 52, 58, 60, 61, 63, 65–67], some researchers have explored using neural networks to generate structurally diverse LiDAR scenes. [5] is the first to explore LiDAR scene generation using generative models (VAE [18] and GAN [13]), exhibiting promising results. Inspired by this, some works use DDPMs [17] to generate 3D scenes [36, 46, 55], achieving superior performance.

Although existing methods have achieved promising LiDAR scene generation results, insufficient training priors often lead to overly smooth and homogeneous scenes, limiting applicability. In this paper, we propose a Self-Conditional Representation Guidance, encouraging DN to learn geometric representations by regularization, improving object details and structural fidelity in generated scenes.

Text-Guided Generation. Natural language can intuitively describe scene content, providing flexible semantic guidance. Benefiting from the available and high-quality Text-Image sample pairs [4, 50], many methods can successfully generate semantically aligned and diverse high-fidelity images from given natural language descriptions [45, 47, 48]. Inspired by these advances, some researchers have explored using text guidance to generate 3D data. Due to the lack of high-quality Text-Point Cloud data, early methods leverage Text-Image priors to bridge the gap between text and point clouds, achieving object-level Text-to-Point Cloud generation [27, 37, 41]. Subsequently, several works attempted to directly generate object-level 3D data from text prompts [33, 57]. Recently, some methods have showed the promise of Text-to-LiDAR scene generation [55].

Some explorations have demonstrated the potential of text-guided LiDAR generation, but the lack of high-quality Text-LiDAR pairs hinder this progress. In this paper, we construct a content-composable Text-LiDAR benchmark and provide a controllability evaluation metric. This relies only on 3D box priors, enabling easy extension to detection datasets and promoting Text-to-LiDAR generation research.

Level	Type	Prompt Example	Sample Distribution	FSID ↓	FPVD ↓	TBR(%) ↑
(■) Object	Quantity	(♠) Two cars.	857,1873,2099,29320	67.32	66.34	31.12
		(♠) There are two cars in the scene.	857,1873,2099,29320	67.54	66.78	30.55
		(♠) Two cars. One car is in front. One car is behind.	857,1873,2099,29320	67.63	66.81	29.88
		(♣) Two cars. → There are two cars in the scene.	857,1873,2099,29320	68.55	67.10	29.23
		(♣) There are two cars in the scene. → Two cars.	857,1873,2099,29320	68.32	67.01	29.94
	Adjusted Text	(♦) Less/More than five cars.	10692,23457	65.10	64.15	60.35
	Location	One car is behind to the right of one pedestrian.	681,478,...,339,468	68.12	66.95	12.23
	Adjusted Text	(♦) No car./One car is around one pedestrian/barrier/truck.	12273,11534,3819,6523	66.74	65.54	23.42
	Orientation	One car is facing backward.	7416,6662,9994,8711,1366	66.45	65.12	37.12
	Adjusted Text	(♦) No car./One car is facing right/left.	1366,16127,16656	65.54	64.32	59.42
(▲) Scene	Weather	Rainy.	6670,27479	65.14	64.55	-
	Time	Night.	3987,30162	65.53	64.74	-
-	Wea., Loc.	(★) Rainy. One car is around one pedestrian	10101,9876,2966,4536,...	66.93	65.84	23.44

Table 1. Results of different text forms. "Text1. → Text2." means that the model is trained with the text form of "Text2", while using "Text1" as conditional input in inference. "Wea., Loc." denotes "Weather, Location". In nuScenes, original descriptions like "***Turn right at intersection, cross bridge, many peds***" are unnatural (the comparison between the original and re-annotated text descriptions in SM).

3. Text Prompt for LiDAR Generation

In this section, we use the re-annotated Text-LiDAR benchmark to evaluate the effect of different text prompts for LiDAR generation quality and controllability, providing the optimized prompt form and scene description insights.

3.1. Annotation and evaluation

Text Annotation. We re-annotated all LiDAR data from 34149 samples in nuScenes with object-level (the target object: "car") and scene-level text descriptions (the annotation process in the supplementary material (SM)). Meanwhile, they are stored independently for flexible text combinations.

Evaluation Metric. We train T2LDM on different text forms, using FID [46] and TBR to evaluate generation quality and controllability. TBR means the matching **R**ate between the **T**ext prompt and the **B**oxes obtained by applying a detector [31] on 10,000 generated samples (details in SM).

3.2. Text Prompt Comparison

Quantity, Location, and Orientation (■). Tab. 1 shows the comparison of generation quality and controllability across different object-level text prompts. Surprisingly, the explicit location prompt yields the worst results. Moreover, the generation controllability is substantially lower than that of other text prompt forms.

Weather and Time (▲). Meanwhile, we also exhibit the results of scene-level text descriptions in Tab. 1, significantly outperforming object-level text prompts in generation.

Text Length (♠). We further evaluate the generation results for different text lengths. In Tab. 1, longer text prompts with similar semantics cause a slight degradation in results. This is because, ***redundant information in longer prompts may hinder the model from capturing key semantics [32].***

Form Transfer (♣). Furthermore, as shown in Tab. 1, text prompts with similar semantics but different forms cause only a slight decrease in results. We believe that ***this benefits from the text encoder [44] effectively identifying semantics and producing reliable features.***

3.3. Text Prompt Analysis

In general, layout-aware text prompts are intuitively expected to enhance generation quality and controllability [68, 69]. However, the results in Tab. 1 indicate otherwise. In fact, this phenomenon can be explained from the perspective of ***sample distribution***. The more dispersed sample distribution can produce richer text descriptions but may cause insufficient training priors due to sample scarcity. This becomes particularly severe when the training data are inadequate. This also provides an explanation: ***the more complex texts typically lead to poorer generation results [30, 40].***

Meanwhile, the above results provide some insights:

- Text prompts should be clear and concise while retaining sufficient semantic information (see **Text Length**).
- A strong semantic-aware text encoder is crucial for text prompt generalization [48] (see **Form Transfer**).
- Annotating Text-LiDAR sample pairs should account for the sample distribution of the dataset to generate appropriate text descriptions for each scene (see Sec. 3.3).

3.4. Text Prompt Optimization

Adjusting Text (♦). Based on the above insights, we adjust the description forms of quantity, position, and orientation in Tab. 1. The improvement in generation quality and controllability shows the reliability of annotating text descriptions from the sample distribution perspective.

Prompt Template (★). Based on the sample distribution and text diversity, we consider the benchmark prompt as "weather, location", since this covers ***the weather, object number, and object layout*** of scenes. We also provide detailed sample distributions of text combinations in SM.

4. Methodology

4.1. Generation Process

Input Representation. Range Map (RM) represents the entire LiDAR scene through a spherical projection of 3D coordinates [35]. The columns and rows represent the LiDAR **H**orizontal (0°-360°, HFoV) and **V**ertical (f_{down}-f_{up},

VFoV) **F**ields **o**f **V**iew. The projection of $\boldsymbol{p}_i = (x, y, z)$ is:

$$u = \frac{1}{2}[1 - arctan(y, x)\pi^{-1}]W,$$
$$v = [1 - (arcsin(zr^{-1}) + f_{up})f^{-1}]H, \quad (1)$$

where (u, v) and (H, W) denote the 2D coordinates and the height, width of RM. Meanwhile, $r = ||\boldsymbol{p}_i||^2$ represents the depth distance of each point $\boldsymbol{p}_i$ from the LiDAR sensor.

Range Map and LiDAR data exhibit a (partially) invertible relationship (LiDAR $\rightarrow$ RM $\rightarrow$ LiDAR), thus can be used for generation tasks. Meanwhile, we use depth r and intensity I as pixel values of RM$\in \mathbb{R}^{H\times W\times 2}$ [36, 55].

Conditional DDPMs for Text-to-LiDAR Generation. Given a Range Map $\boldsymbol{x_0} \sim \mathcal{P}_{RM}$ projected from LiDAR coordinates by Eq. 1, a conditional text $\boldsymbol{c} \sim \mathcal{P}_{text}$, and a prior noise $\boldsymbol{x_T} \sim \mathcal{P}_{noise}$, conditional DDPMs achieve the distribution transformation process between $\mathcal{P}_{RM}$ and $\mathcal{P}_{noise}$ via: a predefined forward process q that gradually adds perturbation to $\boldsymbol{x_0}$ until $\boldsymbol{x_T}$, and a trainable reverse process p_θ that slowly removes noise $\boldsymbol{x_T}$ back to $\boldsymbol{x'_0}$ conditioned on $\boldsymbol{c}$. Meanwhile, the timestep $t \sim \mathcal{U}[1024]$ governs the transition dynamics. In this process, ***to effectively learn the distribution transformation, conditional DDPMs typically require sufficient training priors to match*** $\mathcal{P}_{RM}$ ***[54, 70].***

Then, the training objective of conditional DDPMs is:

$$L(\theta) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}||\boldsymbol{v} - v_\theta(\boldsymbol{x_t}, t, \boldsymbol{c})||^2, \quad (2)$$

where the target $\boldsymbol{v}$ can be converted into $\boldsymbol{\epsilon}$ or $\boldsymbol{x_0}$ (the derivation in SM). Meanwhile, unconditional generation can be regarded as a special case of conditional generation ($\boldsymbol{c} = \emptyset$) [42]. Therefore, in this paper, we can achieve the classifier-free guidance (CFG) [16] by alternately training unconditional DDPMs and conditional DDPMs.

Subsequently, we can iteratively transform $\boldsymbol{x_T}$ sampled from $\mathcal{P}_{noise}$ to $\boldsymbol{x'_0} \sim \mathcal{P}_{RM}$ by the trained v_θ in inference.

Finally, $\boldsymbol{x'_0}$ is converted back to the 3D coordinates to generate the LiDAR scene using the inverse of Eq. 1 [35].

4.2. Self-Conditioned Representation Guidance

Unlike easily accessible Text-Image data [23, 62], Text-LiDAR pairs are scarce due to costly collection and annotation [5, 36, 55, 71]. This often leads to insufficient training priors for generative models, resulting in overly smooth results that lack detailed objects in LiDAR scenes (see Fig. 1). In image generation, some methods leverage pretrained priors to enhance the representation capacity of generative models, achieving promising generation performance [23, 62]. However, there are some limitations:

- Requiring large-scale pretrained knowledge priors [39].
- Involving more costly multi-stage training.

In this paper, we propose a Self-Conditional Representation Guidance (SCRG) that employs a Guidance Network

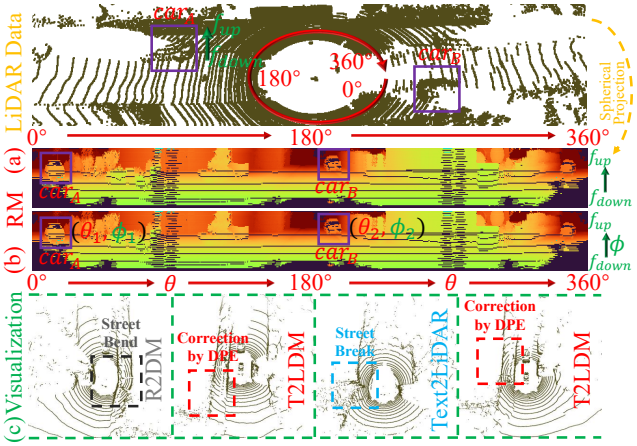

Figure 2. (a) In LiDAR space, car_A is at ***the front-right*** of car_B, but it appears as ***the left*** in RM (window shift). (b) By defining the horizontal angle θ and vertical angle ϕ in RM, DPE provides true directional priors, enabling the model to correctly perceive object orientations in the scene. For example, the model can clearly understand the relative position between $car_A(\theta_1,\phi_1)$ and $car_B(\theta_2,\phi_2)$. (c) Existing methods produce bend or broken streets due to ***directional confusion***. T2LDM generates realistic ones.

(GN, x_ϕ) to learn geometric features with reconstruction details from data distribution in an end-to-end manner. GN can provide adaptive perturbation and condition supervision signals for the Denoising Network (DN, v_θ) to effectively learn geometric details, while detached during inference.

Specifically, GN receives the multi-level perturbation features $F^{v_\theta}_{noise}$ with conditional guidance from DN, aligning real coordinates ($\boldsymbol{x_0}$) to reconstruct geometric details:

$$L(\phi) = ||\boldsymbol{x_0} - \boldsymbol{x_\phi}(\boldsymbol{x_0}, F^{v_\theta}_{noise})||^2. \quad (3)$$

Subsequently, to inject regularization for DN, $F^{v_\theta}_{noise}$ is aligned with the multi-scale reconstruction features $F^{x_\phi}_{recon}$ from GN (see Fig. 3 and Fig. 9(bottom)):

$$L_{SCRG} = l_{recon}(F^{x_\phi}_{recon} - F^{v_\theta}_{noise}), \quad (4)$$

where $l_{recon}(\cdot)$ is a reconstruction loss (the cosine similarity in this paper, the additional ablation studies in SM).

This simple and effective approach:

- **With Lower Training Cost.** GN participates in gradient backpropagation only in the early stage and provides adaptive regularization to DN in an end-to-end manner.
- **Without Inference Cost.** The detachable design of GN prevents cost and information leakage in inference.
- **With Faster Convergence.** The regularization from GN guides DN to learn high-frequency semantics for faster early-stage convergence (see Tab. 8 and Fig. 9(bottom)).

4.3. Directional Position Encoding

RM represents the entire LiDAR scene by flattening the spherical projection [35]. However, window-based opera-

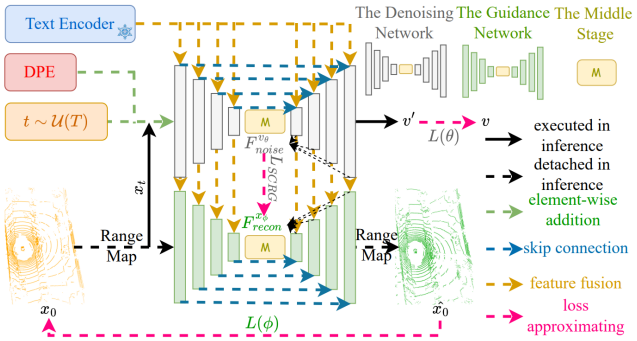

Figure 3. The overall framework of T2LDM. The Text Encoder (TE) encodes text prompts to generate semantically reliable features. Meanwhile, the Denoising Network (DN) models the denoising process under text guidance, DPE, and timestep. Furthermore, the Guidance Network (GN) introduces regularization with reconstruction details for DN while detached during inference.

tions (e.g., convolution or local attention) perceive RM as a rectangular image rather than a circular image. This often leads to ***directional confusion***, making the model struggle to understand proper object orientations in the scene. The effect is most evident in distorted streets, as the starting angle is typically defined at the center of the street (see Fig. 2).

In this paper, we design a Directional Position Encoding (DPE) for RM. By encoding the HFoV and VFoV angles, DPE can inject spherical geometric orientation priors, making the model perceive the true position of content in RM.

Specifically, given $\boldsymbol{x} \in \mathbb{R}^{b\times c\times h\times w}$ from v_θ or x_ϕ, DPE first defines the angle coordinate of each pixel (i, j) in RM:

$$\begin{aligned} \boldsymbol{\theta} &= 2\pi - (2\pi - 0) * (i + 0.5)/w, \\ \boldsymbol{\phi} &= f_{up} - (f_{up} - f_{down}) * (j + 0.5)/h. \end{aligned} \tag{5}$$

Then, the Fourier expansion and a learnable gating are applied for the angle coordinate $(\boldsymbol{\theta}, \boldsymbol{\phi})$:

$$\begin{aligned} \mathrm{DPE}(\boldsymbol{\theta}, \boldsymbol{\phi}) &= Fourier^K(\boldsymbol{\theta}, \boldsymbol{\phi}), \\ \boldsymbol{x'} &= \boldsymbol{x} + \alpha * \mathrm{DPE}(\boldsymbol{\theta}, \boldsymbol{\phi}), \end{aligned} \tag{6}$$

where α means a learnable gating parameter. K denotes the number of Fourier expansion terms, $Fourier^K(\boldsymbol{\theta}, \boldsymbol{\phi}) = \bigoplus_{k=0}^{K-1}[\sin(2^k\boldsymbol{\theta}), \cos(2^k\boldsymbol{\theta}), \sin(2^k\boldsymbol{\phi}), \cos(2^k\boldsymbol{\phi})]$.

DPE encodes pixel angles via multi-level Fourier expansions to provide multi-scale directional priors in RM. Meanwhile, the learnable gating adaptively modulates the weights of multi-frequency features.

4.4. Network Architecture

In this section, we present T2LDM overall architecture, consisting of three key components in Fig. 3: the Text Encoder (TE), the Denoising Network (DN), and the Guidance Network (GN) (parameters and optimizations in SM).

The Text Encoder. TE encodes text prompts to provide semantic conditional features. The text prompt generalization largely depends on TE (see Sec. 3.3). Therefore, we use the frozen CLIP to produce 768-dimensional text features with Text-to-Image semantic alignment [27, 41].

The Denoising Network. DN models the denoising, determining the generation results, following the U-Net architecture [47]. Each stage consists of Attention Block (AB) and Residual Block (RB) in the encoder and decoder.

Specifically, AB receives text features F^{CLIP}_{text} from TE as conditional guidance. This fuses the projected features $F^{v_\theta}_{noise} \in \mathbb{R}^{l\times C^{v_\theta}} \rightarrow (Q) \in \mathbb{R}^{l\times C}$ and $F^{CLIP}_{text} \in \mathbb{R}^{n\times 768} \rightarrow (K, V) \in \mathbb{R}^{n\times C}$ by a cross-attention block:

$$\begin{aligned} O &= mlp(WV) + F^{v_\theta}_{noise}, \\ F &= ffn(O) + O, \end{aligned} \tag{7}$$

where $W \in \mathbb{R}^{l\times n} = softmax(\frac{QK^T}{\sqrt{C}})$ and $l = h \times w$.

Replacing F^{CLIP}_{text} with $F^{v_\theta}_{noise}$ means unconditional generation. The timestep t and DPE are introduced into RB to identify denoising level and enhance scene fidelity.

The Guidance Network. GN produce the supervision signals with reconstruction details for DN (see Sec. 4.2). This follows DN architecture with four stages in the encoder and decoder. To provide perturbation and conditional adaptation regularization, GN receives noise features from DN by Eq. 7. Meanwhile, we use only RM as input to ensure GN focuses on learning geometric features of data distribution.

4.5. Training and Inference

Training. As mentioned earlier (see Sec. 4.1 and Sec. 4.2), T2LDM models the denoising process with SCRG. Therefore, the training objective is:

$$L_{total} = L(\theta) + L(\phi) + \lambda L_{SCRG}, \tag{8}$$

where λ means an epoch-wise weighting factor (details in SM). Meanwhile, x_ϕ only participates in gradient backpropagation for the first 100K steps, then remains frozen.

Inference. T2LDM iteratively transforms $\boldsymbol{x_T}$ into $\boldsymbol{x'_0}$ by only v_θ, due to the detachable design of x_ϕ (see Sec. 4.2):

$$\boldsymbol{x_{t-1}} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x_t} - \frac{1-\alpha_t}{\sigma_t}[\sigma_t\boldsymbol{x_t} + \sqrt{\bar{\alpha}_t}\, v_\theta]) + \tilde{\sigma}_t\boldsymbol{\epsilon}, \tag{9}$$

where $\sigma_t = \sqrt{1-\bar{\alpha}_t}$, $\tilde{\sigma}_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)}$.

5. Experiments

5.1. Experiment Setup

Dataset. Two LiDAR benchmarks are used for training and evaluation: nuScenes [6] (32-beam, 34,149 samples) and KITTI-360 [26] (64-beam, 76,165 samples). Meanwhile, the LiDAR data are projected into $RM_{32beam} \in \mathbb{R}^{32\times 1024\times 2}$ and $RM_{64beam} \in \mathbb{R}^{64\times 1024\times 2}$ (see Sec. 4.1).

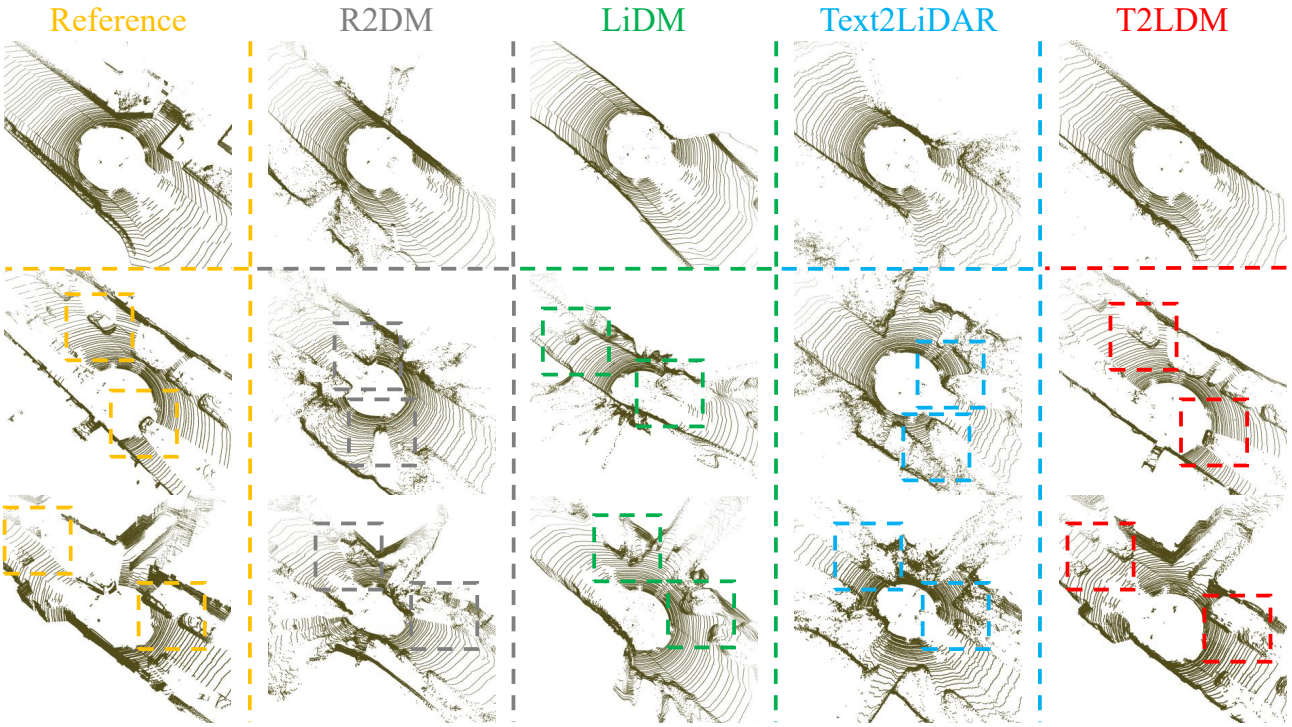

Figure 4. The generated visualization results on KITTI-360. Due to insufficient training priors, existing methods can only generate high-quality scenes with a few objects (top row). In contrast, T2LDM produces fine-grained geometric details even in complex multi-object scenes (bottom row). This is crucial for models to recognize 3D scenes in downstream tasks. For more visualizations, please refer to SM.

Metric. FID (FSVD, FPVD), JSD, and MMD ($\times 10^{-4}$) are used for generation quality evaluation [46]. For fair comparison, we compute the true distribution over ***all real samples*** instead of randomly selecting subsets for FID [46, 55]. Meanwhile, for text-guided controllability, we propose TBR (see Sec. 3.1), the matching rate between text semantics and 3D boxes obtained by a detector [31].

5.2. Unconditional Generation

64-Beam LiDAR. We first evaluate the generation quality on KITTI-360. Tab. 2 shows that T2LDM better matches the true distribution, as evidenced by the significantly lower FID. Benefiting from SCRG, T2LDM can more effectively learn realistic scene details from the data distribution. Therefore, compared with methods lacking sufficient prior training, T2LDM produces results with finer geometric details. Fig. 4 further presents the qualitative results.

Methods	Gen. Sam.	Rea. Sam.	FSVD↓	FPVD↓	JSD↓	MMD↓
LiDARVAE [5]	10000	76165	281.14	286.14	0.35	6.84
LiDARGAN [5]	10000	76165	346.23	339.55	0.38	5.43
ProjectedGAN [49]	10000	76165	187.89	201.62	0.33	3.45
LiDARGen [71]	10000	76165	238.72	243.69	0.32	3.93
LiDM [46]	10000	76165	211.68	230.19	0.35	4.78
R2DM [36]	10000	76165	31.82	35.94	0.32	4.05
Text2LiDAR [55]	10000	76165	51.55	54.82	0.33	4.11
T2LDM	10000	76165	21.12	25.39	0.30	3.35

Table 2. The results on KITTI-360. T2LDM significantly outperforms existing methods on all metrics.

32-Beam LiDAR. We also evaluate on the 32-beam benchmark. Compared with KITTI-360, nuScenes with fewer points and the larger spatial distance between points is more challenging due to the harder-to-capture geometric details. This also leads to poor performance for existing methods on nuScenes. However, T2LDM can achieve excellent generation results in Tab. 3. As described in Sec. 4.2 and Sec. 4.3, T2LDM captures effective geometric and directional priors during training through SCRG and DPE, enhancing the ability to perceive geometric details from the data distribution. Fig. 5 and Fig. 1 further demonstrates T2LDM can effectively generate rich and diverse details in sparse scenes.

Methods	Gen. Sam.	Rea. Sam.	FSVD↓	FPVD↓	JSD↓	MMD↓
R2DM [36]	10000	34149	86.54	83.97	0.42	5.02
Text2LiDAR [55]	10000	34149	85.98	80.94	0.34	3.45
T2LDM	10000	34149	64.21	62.85	0.26	3.01

Table 3. The results on nuScenes. T2LDM achieves superior generation results across all metrics in sparse scenes.

5.3. Text-Guided Generation

Unlike other conditions with various constraints, text prompts are more accessible for human beings and can provide customized and diverse scene descriptions.

We further validate the results on Text-to-LiDAR generation for T2LDM. Benefiting from detection priors, we can measure the generation controllability using the matching rate (TBK) between text prompts and 3D boxes obtained from a detector [31]. Meanwhile, as described in Sec. 3.4, we perform evaluations on "weather, location". Tab. 4 presents that T2LDM exhibits remarkable results in

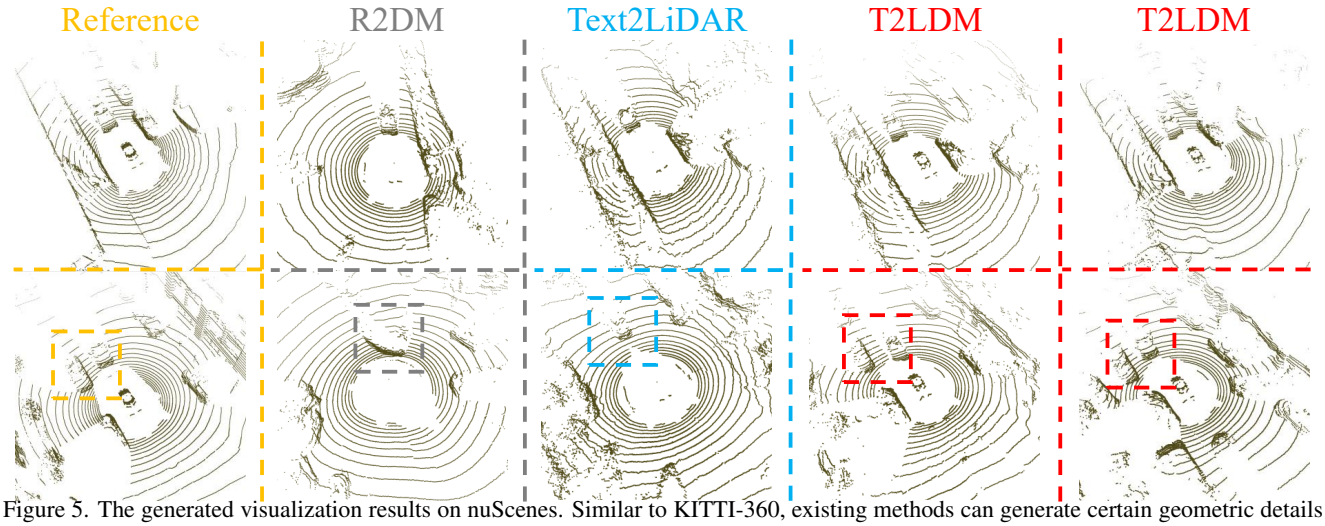

Figure 5. The generated visualization results on nuScenes. Similar to KITTI-360, existing methods can generate certain geometric details in scenes with few objects (top row) but struggle to handle complex multi-object scenes (bottom row), duo to the sufficient training data. This becomes more pronounced in the sparse scenes of nuScenes. In comparison, T2LDM can generate detailed objects even in multi-object scenes. Fig. 1 also shows that T2LDM can generate diverse structures for the same scene. More visualizations are provided in SM.

generation quality and controllability. With the perturbation and condition-guided adaptation regularization from GN, T2LDM can further perceive conditional features, improving the understanding of guided information. Furthermore, Fig. 8 shows the qualitative comparison.

Methods	Gen. Sam.	Rea. Sam.	FSVD↓	FPVD↓	JSD↓	MMD↓	TBK(%)↑
R2DM [36]	10000	34149	91.15	88.55	0.45	5.11	15.45
Text2LiDAR [55]	10000	34149	90.13	87.62	0.38	4.01	17.15
T2LDM	10000	34149	66.93	65.84	0.28	3.05	23.44

Table 4. The text-guided results on nuScenes. T2LDM exhibits outstanding performance in generation quality and controllability.

5.4. Other Conditional Generation

By freezing unconditional DN, T2LDM can achieve various conditional tasks. ***This also marks the first exploration of ControlNet [64] into 3D generation in non-latent DDPMs*** (please refer to the implementation details in SM).

Methods	4×			8×		
	CD↓	MSE↓	EMD↓	CD↓	MSE↓	EMD↓
Grad-PU [15]	0.400	4.169	2.324	0.364	4.031	2.142
PUDM [42]	0.198	4.275	2.124	0.103	4.102	1.914
T2LDM	0.104	3.610	1.987	0.074	3.574	1.910

Table 5. The results of the $4\times$ rate and the $8\times$ rate on nuScenes. T2LDM exhibits significantly upsampling results.

Sparse-to-Dense Generation. We downsample the training set (28,140 samples) by $4\times$ using FPS as sparse LiDAR for training, while the original and $2\times$ upsampled validation sets (6019 samples) are used as $4\times$ and $8\times$ Ground Truth for evaluation. We follow existing methods [15, 42] by directly validating the PU-GAN [22] pretrained model on nuScenes. Meanwhile, we find that PUDM shows better qualitative but inconsistent quantitative results than Grad-PU. For fair comparison, we normalize point coordinates to [0,1] (CD$\times 10^{-5}$, MSE$\times 10^{-5}$, EMD$\times 10^{-3}$). Tab. 5 presents the remarkable upsampling results for T2LDM. Fig. 6(a) further illustrates the superior qualitative results.

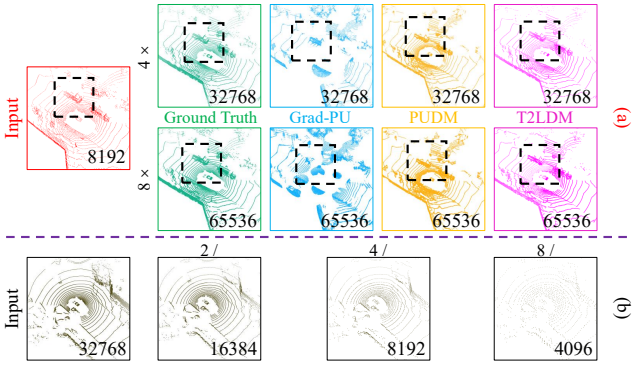

Figure 6. (a) Qualitative results of upsampling. (b) Without retraining, T2LDM can perform downsampling at arbitrary rates.

Dense-to-Sparse Generation. Meanwhile, since the output LiDAR data shape is determined by the input noise size, we can directly achieve downsampling by upsampling-enabled T2LDM. Fig. 6(b) shows the Dense-to-Sparse results.

Semantic-to-LiDAR Generation. Furthermore, we also implement Semantic-to-LiDAR generation on nuScenes. Tab. 6 and Fig. 7 present the quantitative and qualitative results, respectively (SemanticKITTI [2] results in SM).

Methods	Gen. Sam.	Rea. Sam.	FSVD↓	FPVD↓	JSD↓	MMD↓
T2LDM+Uncon.	10000	34149	64.21	62.85	0.26	3.01
T2LDM+Seman.	10000	34149	62.91	60.54	0.23	2.94

Table 6. The Semantic-to-LiDAR results on nuScenes. T2LDM achieves excellent results for semantic map guidance generation.

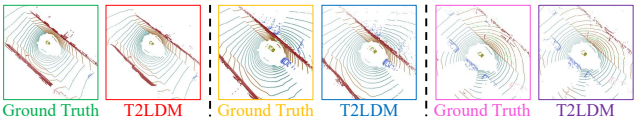

Figure 7. Semantic-to-Lidar results on nuScenes. With the DN frozen, T2LDM shows remarkable semantic-guided generation.

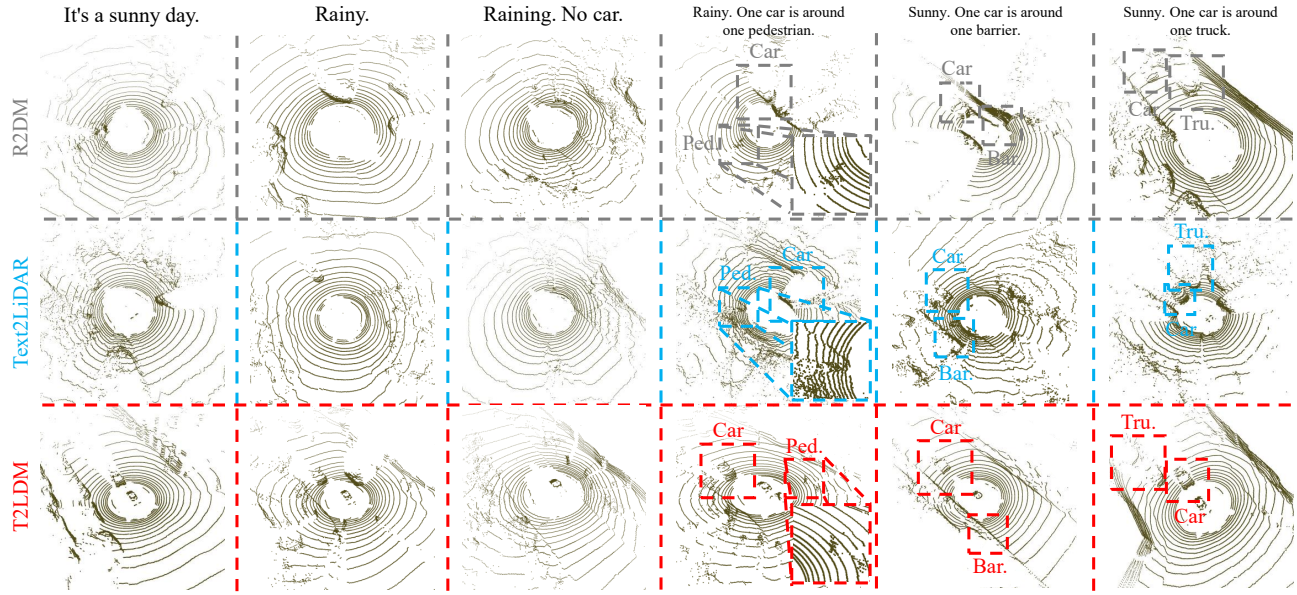

Figure 8. Text-guided generation results on nuScenes. Existing methods produce overly smooth results with insufficient object details, making the difficult to satisfy the text semantics. However, T2LDM shows superior detail generation that aligns well with the text prompts.

5.5. Ablation Study

Component Effectiveness. We first perform ablations for component effectiveness. Tab. 7 shows that removing SCRG and DPE leads to a significant drop for the generation quality and controllability of T2LDM. As discussed in Sec. 4.2 and Sec. 4.3, insufficient training priors make learning real details from the data distribution difficult for the model, leading to blurry objects in the generated scenes.

Methods	Gen. Sam.	Rea. Sam.	FSVD↓	FPVD↓	JSD↓	MMD↓	TBK(%)↑
T2LDM$^{\emptyset}$	10000	34149	73.64	71.91	0.34	3.21	19.32
T2LDM^{D}	10000	34149	71.32	70.44	0.32	3.15	20.95
T2LDM^{S}	10000	34149	68.45	67.77	0.30	3.07	22.15
T2LDM	10000	34149	66.93	65.84	0.28	3.05	23.44

Table 7. Ablation study of component effectiveness for text-guided generation on nuScenes. T2LDM$^{\emptyset}$, T2LDM^{D}, and T2LDM^{S} denote removing DPE and SCRG, keeping only DPE, and keeping only SCRG, respectively. DPE and SCRG can provide effective priors and regularization, enhancing scene fidelity.

Convergence Speed. We further evaluate the effect of SCRG on the convergence speed for T2LDM. Fig. 9(top) shows that SCRG allows DN to capture high frequency details early in training, generating detailed scene structures. Fig. 9(bottom) presents that GN can learn rich geometric detail features to offer effective regularization. Fig. 9(right) shows that SCRG enables faster and more stable convergence of T2LDM. Tab. 8 presents results at 30k iterations.

Methods (30k Itera.)	Inf. Param.	Inf. Steps	Gen. Sam.	FSVD↓	FPVD↓	JSD↓	MMD↓
R2DM [36]	31.1M	1024	10000	175.82	152.57	0.55	10.08
Text2LiDAR [55]	45.8M	1024	10000	340.95	320.43	0.84	16.61
T2LDM^{D}	30.4M	1024	10000	91.32	88.44	0.45	1.45
T2LDM	30.4M	1024	10000	47.29	55.57	0.35	0.55

Table 8. The results on KITTI-360 at 30k iterations. SCRG enables T2LDM to learn high-frequency semantics early.

End-to-End vs. Pretrained Mode. We also conduct the ablation study on the end-to-end and pretrained paradigms of SCRG. As shown in Tab. 9, pretrained training even leads to performance degradation, since GN cannot perceive DN features to provide adaptive regularization [21]. In contrast, end-to-end paradigm enables joint training of GN and DN, allowing GN to produce feature-aware supervision signals.

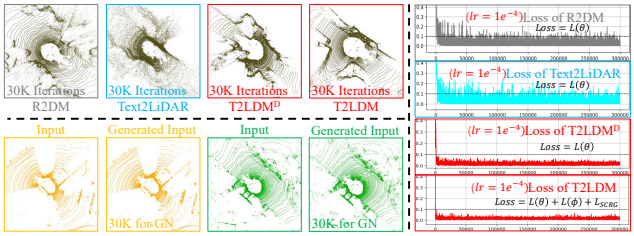

Figure 9. Ablation study of convergence speed on KITTI360. The loss combination of T2LDM shows more stable and superior.

Methods	Gen. Sam.	Rea. Sam.	FSVD↓	FPVD↓	JSD↓	MMD↓
T2LDM^{D}	10000	34149	68.11	67.32	0.31	3.11
Pretrained Mode	10000	34149	68.77	67.94	0.33	3.12
End-to-End Mode	10000	34149	64.21	62.85	0.26	3.01

Table 9. Ablation study of end-to-end vs. pretrained training for SCRG on nuScenes. The end-to-end mode yields better results.

6. Conclusion

In this paper, we proposed a Text-to-LiDAR diffusion model that leverages a self-conditioned representation guidance to enhance details in generated LiDAR scenes. Meanwhile, a directional position prior is used to resolve directional confusion, correcting road distortion. Moreover, we design a 3D box-based annotation scheme to construct a content-composable Text-LiDAR benchmark, offering a controllability generation metric and insights to encourage researchers to focus on Text-to-LiDAR generation.

Acknowledgments

This work was supported in part by the Frontier Technologies R&D Program of Jiangsu under grant BF2024070, in part by the National Natural Science Foundation of China under Grant 62471235, in part by Inspur Storage Qinglan Foundation and Shandong Information Storage System Technology Innovation Center, in part by Hunan Natural Science Foundation Project (No. 2025JJ50338) and Shanghai Education Committee AI Project (No. JWAIYB-2), in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX25_0754, and in part by PNRR FAIR - Future AI Research (PE00000013).

References

[1] Xiaoqi An, Lin Zhao, Chen Gong, Jun Li, and Jian Yang. Pre-training a density-aware pose transformer for robust lidar-based 3d human pose estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1755–1763, 2025. 1

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 7

[3] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020. 1

[4] Kakao Brain. Coyo-700m: Large-scale image-text pairs dataset. https://github.com/kakaobrain/coyo-dataset, 2023. Accessed: 2025-10-22. 1, 2

[5] Lucas Caccia, Herke Van Hoof, Aaron Courville, and Joelle Pineau. Deep generative modeling of lidar data. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5034–5040. IEEE, 2019. 2, 4, 6

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 5

[7] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12929–12938, 2021. 2

[8] Ping Chen, Xingpeng Zhang, Zhaoxiang Liu, Huan Hu, Xiang Liu, Kai Wang, Min Wang, Yanlin Qian, and Shiguo Lian. Optimizing for the shortest path in denoising diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18021–18030, 2025.

[9] Ping Chen, Zezhou Chen, Xingpeng Zhang, Yanlin Qian, Huan Hu, Xiang Liu, Zipeng Wang, Xin Wang, Zhaoxiang Liu, Kai Wang, et al. Beyond geometry: Artistic disparity synthesis for immersive 2d-to-3d. *arXiv preprint arXiv:2603.05906*, 2026.

[10] Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-ntfm: A unified foundation model for eeg signal representation learning. *arXiv preprint arXiv:2509.24222*, 2025.

[11] Chen Feng and Ioannis Patras. MaskCon: Masked Contrastive Learning for Coarse-Labelled Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025. 2

[13] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[14] Martin Hahner, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15283–15292, 2021. 2

[15] Yun He, Danhang Tang, Yinda Zhang, Xiangyang Xue, and Yanwei Fu. Grad-pu: Arbitrary-scale point cloud upsampling via gradient descent with learned distance functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2023. 7

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] Qizhen Lan and Qing Tian. Instance, scale, and teacher adaptive knowledge distillation for visual detection in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 8(3):2358–2370, 2022. 2

[20] Qizhen Lan and Qing Tian. Acam-kd: adaptive and cooperative attention masking for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3957–3966, 2025. 2

[21] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025. 8

[22] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7203–7212, 2019. 7

[23] Tianhong Li, Dina Katabi, and Kaiming He. Self-conditioned image generation via generating representations. *CoRR*, 2023. 2, 4

[24] Yuqi Li, Yanli Li, Kai Zhang, Fuyuan Zhang, Chuanguang Yang, Zhongliang Guo, Weiping Ding, and Tingwen Huang. Achieving fair medical image segmentation in foundation models with adversarial visual prompt tuning. *Information Sciences*, page 122501, 2025. 2

[25] Yuqi Li, Chuanguang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2025. 2

[26] Yiyi Liao, Jun Xie, and Andreas Geiger. 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. 5

[27] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023. 2, 5

[28] Jiuming Liu, Guangming Wang, Weicai Ye, Chaokang Jiang, Jinru Han, Zhe Liu, Guofeng Zhang, Dalong Du, and Hesheng Wang. Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15109–15119, 2024. 2

[29] Jiuming Liu, Weicai Ye, Guangming Wang, Chaokang Jiang, Lei Pan, Jinru Han, Zhe Liu, Guofeng Zhang, and Hesheng Wang. Difflow3d: Hierarchical diffusion models for uncertainty-aware 3d scene flow estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2025. 2

[30] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pages 423–439. Springer, 2022. 3

[31] Shuai Liu, Mingyue Cui, Boyang Li, Quanmin Liang, Tinghe Hong, Kai Huang, and Yunxiao Shan. Fshnet: Fully sparse hybrid network for 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8900–8909, 2025. 3, 6

[32] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–23, 2022. 1, 3

[33] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2837–2845, 2021. 2

[34] Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Lidarsim: Realistic lidar simulation by leveraging the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11167–11176, 2020. 2

[35] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4213–4220. IEEE, 2019. 3, 4

[36] Kazuto Nakashima and Ryo Kurazume. Lidar data synthesis with denoising diffusion probabilistic models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14724–14731. IEEE, 2024. 1, 2, 4, 6, 7, 8

[37] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2

[38] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. Prompting ai art: An investigation into the creative skill of prompt engineering. *International journal of human–computer interaction*, 41(16):10207–10229, 2025. 1

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 5

[42] Wentao Qu, Yuantian Shao, Lingwu Meng, Xiaoshui Huang, and Liang Xiao. A conditional denoising diffusion probabilistic model for point cloud upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20786–20795, 2024. 4, 7

[43] Wentao Qu, Jing Wang, YongShun Gong, Xiaoshui Huang, and Liang Xiao. An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27325–27335, 2025. 1

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 3

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 2

[46] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14738–14748, 2024. 2, 3, 6

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 5

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2, 3

[49] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 6

[50] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1, 2

[51] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025. 2

[52] Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[53] Sven Teufel, Georg Volk, Alexander Von Bernuth, and Oliver Bringmann. Simulating realistic rain, snow, and fog variations for comprehensive performance characterization of lidar perception. In *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, pages 1–7. IEEE, 2022. 2

[54] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36:72137–72154, 2023. 4

[55] Yang Wu, Kaihua Zhang, Jianjun Qian, Jin Xie, and Jian Yang. Text2lidar: Text-guided lidar point cloud generation via equirectangular transformer. In *European Conference on Computer Vision*, pages 291–310. Springer, 2024. 1, 2, 4, 6, 7, 8

[56] Yang Wu, Yun Zhu, Kaihua Zhang, Jianjun Qian, Jin Xie, and Jian Yang. Weathergen: A unified diverse weather generator for lidar point clouds via spider mamba diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17019–17028, 2025. 2

[57] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8929–8939, 2023. 2

[58] Feng Xu, Guangyao Zhai, Xin Kong, Tingzhong Fu, Daniel FN Gordon, Xueli An, and Benjamin Busam. Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models. *arXiv preprint arXiv:2512.05107*, 2025. 2

[59] Donglin Yang, Xinyu Cai, Zhenfeng Liu, Wentao Jiang, Bo Zhang, Guohang Yan, Xing Gao, Si Liu, and Botian Shi. Realistic rainy weather simulation for lidars in carla simulator. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 951–957. IEEE, 2024. 2

[60] Jianjian Yin, Tao Chen, Gensheng Pei, Huafeng Liu, Yazhou Yao, Liqiang Nie, and Xiansheng Hua. Semi-supervised semantic segmentation with multi-constraint consistency learning. *IEEE TMM*, 27:6449–6461, 2025. 2

[61] Jianjian Yin, Yi Chen, Zhichao Zheng, Junsheng Zhou, and Yanhui Gu. Uncertainty-participation context consistency learning for semi-supervised semantic segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[62] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 2, 4

[63] Hansheng Zeng, Yuqi Li, Ruize Niu, Chuanguang Yang, and Shiping Wen. Enhancing spatiotemporal prediction through the integration of mamba state space models and diffusion transformers. *Knowledge-Based Systems*, 2025. 2

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 7

[65] Ruilin Zhang, Haiyang Zheng, and Hongpeng Wang. Cnmbi: Determining the number of clusters using center pairwise matching and boundary filtering. In *International Conference on Advanced Data Mining and Applications*, pages 262–277. Springer, 2023. 2

[66] Ruilin Zhang, Haiyang Zheng, and Hongpeng Wang. Tdec: Deep embedded image clustering with transformer and distribution information. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 280–288, 2023.

[67] Haiyang Zheng, Ruilin Zhang, and Hongpeng Wang. Deep image clustering based on curriculum learning and density information. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 330–338, 2024. 2

[68] Junwei Zhou, Xueting Li, Lu Qi, and Ming-Hsuan Yang. Layout-your-3d: Controllable and precise 3d generation with 2d blueprint. *arXiv preprint arXiv:2410.15391*, 2024. 3

[69] Yang Zhou, Zongjin He, Qixuan Li, and Chao Wang. Layoutdreamer: Physics-guided layout for text-to-3d compositional scene generation. *arXiv preprint arXiv:2502.01949*, 2025. 3

[70] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Domainstudio: Fine-tuning diffusion models for domain-driven image generation using limited data. *International Journal of Computer Vision*, 133(10):7012–7036, 2025. 4

[71] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point clouds. In *European Conference on Computer Vision*, pages 17–35. Springer, 2022. 2, 4, 6