Fast Training of Diffusion Transformer with Extreme Masking for 3D Point Clouds Generation

Shentong Mo[1] Enze Xie[2] Yue Wu[2] Junsong Chen[2]
Matthias Nießner[3] Zhenguo Li[2]

[1] MBZUAI
[2] Huawei Noah's Ark Lab
[3] TUM

Abstract. Diffusion Transformers have recently shown remarkable effectiveness in generating high-quality 3D point clouds. However, training voxel-based diffusion models for high-resolution 3D voxels remains prohibitively expensive due to the cubic complexity of attention operators, which arises from the additional dimension of voxels. Motivated by the inherent redundancy of 3D compared to 2D, we propose FastDiT-3D, a novel masked diffusion transformer tailored for efficient 3D point cloud generation, which greatly reduces training costs. Specifically, we draw inspiration from masked autoencoders to dynamically operate the denoising process on masked voxelized point clouds. We also propose a novel voxel-aware masking strategy to adaptively aggregate background/foreground information from voxelized point clouds. Our method achieves state-of-the-art performance with an extreme masking ratio of nearly 99%. Moreover, to improve multi-category 3D generation, we introduce Mixture-of-Expert (MoE) in 3D diffusion model. Each category can learn a distinct diffusion path with different experts, relieving gradient conflict. Experimental results on the ShapeNet dataset demonstrate that our method achieves state-of-the-art high-fidelity and diverse 3D point cloud generation performance. Our FastDiT-3D improves 1-Nearest Neighbor Accuracy and Coverage metrics when generating 128-resolution voxel point clouds, using only 6.5% of the original training cost.

Keywords: Diffusion Transformers · Efficient Training · 3D Point Clouds Generation

1 Introduction

Recent breakthroughs in Diffusion Transformers have made remarkable strides in advancing the generation of high-quality 3D point clouds. Notably, the current state-of-the-art (SOTA), DiT-3D [23], leveraged a diffusion transformer architecture for denoising voxelized point clouds, significantly outperformed previous UNet-based methods such as LION [33] by improving 1-Nearest Neighbor Accuracy (1-NNA) at 8.49% and Coverage (COV) at 6.51% in terms of Chamfer

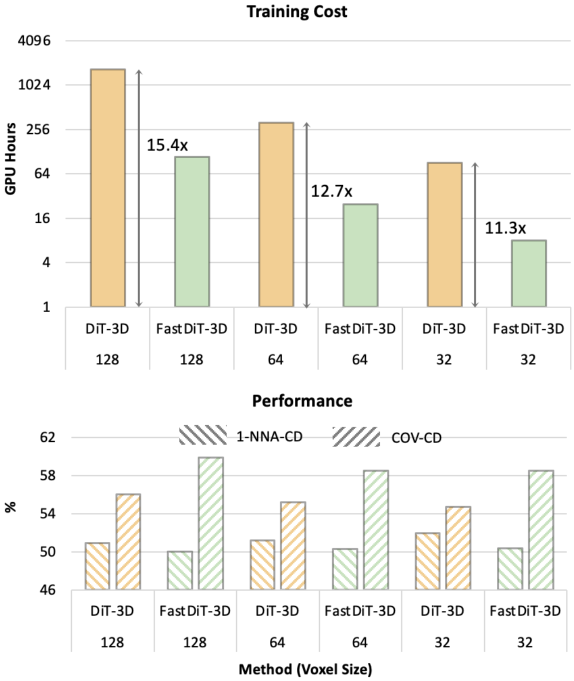

Fig. 1: Comparison of the proposed FastDiT-3D with DiT-3D in terms of different voxel sizes on training costs (lower is better) and COV-CD performance (higher is better). Our method achieves faster training while exhibiting superior performance.

Distance (CD). They also achieved superior performance compared to the previous best UNet-based mesh generation model MeshDiffusion [21]. Based on their excellent experimental results, adopting transformer architecture is expected to be the mainstream approach for 3D shape generation tasks. Despite their efficacy, the voxel-based diffusion transformer's training overhead significantly increases primarily due to the additional dimension when transferring from 2D to 3D. This results in cubic complexity associated with attention mechanisms within the volumetric space. For instance, training voxels of 128 × 128 × 128 takes 1,668 A100 GPU hours. Such a large amount of computational resources is the bottleneck to further increasing the input voxel size or scaling up these model architectures. The training efficiency of diffusion transformers in 3D shape generation is still an unsolved problem.

In image generation and visual recognition, masked training [5, 6, 15, 34] is widely adopted to improve training efficiency, which significantly reduces training time and memory but does not comprise the performance. Considering the high redundancy of 3D voxels, only a partial of the volumetric space is occupied. It is possible to generate high-fidelity 3D shape training on a subset of voxels.

In this work, we introduce FastDiT-3D, a novel diffusion transformer architecture explicitly designed to generate 3D point clouds efficiently. Inspired by masked autoencoders [15], we propose a dynamic denoising operation on selectively masked voxelized point clouds. We further propose a novel foreground-background aware masking strategy, which adaptly aggregates information by differentiating between the information-rich foreground and information-poor background within the point clouds. This innovative approach achieves an outstanding masking ratio, with almost 98% of input voxels masked, superior to the 50% observed in 2D [34], leading to a remarkable 13X acceleration in training speed, as shown in Fig. 1. Moreover, to address the heightened computational demands posed by the increased token length in 3D contexts, we integrate 3D window attention mechanisms within the decoder's Transformer blocks. Our training regimen employs a dual-objective strategy, applying a denoising objective to unmasked patches while masked patches undergo a distinct point cloud generation objective. Our approach not only accelerates the training process but also achieves SOTA performance.

To enhance the capability of point cloud generation across diverse categories, we incorporate Mixture of Expert (MoE) layers within the Transformer blocks. In this way, we transform a dense 3D diffusion model into a sparse one. Each category can learn a distinct diffusion path, and each diffusion path is composed of different experts across different layers. This design greatly alleviates the challenge of difficult gradient optimization caused by multi-category joint training.

Our comprehensive evaluation on the ShapeNet dataset conclusively attests to FastDiT-3D's state-of-the-art performance in generating high-fidelity and diverse 3D point clouds across categories, evidenced by improved 1-NNA and COV metrics for 128-resolution voxel point clouds. Remarkably, our model achieves these results at a mere **6.5%** of the original training cost. Qualitative visualizations further corroborate FastDiT-3D's proficiency in rendering detailed 3D shapes. A series of ablation studies underscore the critical roles played by the foreground-background aware masking, the encoder-decoder architecture, and the dual training objectives in the adept learning of our FastDiT-3D. Lastly, incorporating MoE distinctly showcases the model's effectiveness in accommodating multiple categories through a unified global model.

Our main contributions can be summarized as follows:

- We present a fast diffusion transformer based on encoder-decoder architecture for point cloud shape generation, called FastDiT-3D, that can efficiently perform denoising operations on masked voxelized point clouds with an extreme masking ratio, which masks 99% of the background and 95% of the foreground.
- We propose a novel foreground-background aware masking mechanism to select unmasked patches for efficient encoding and Mixture of Expert (MoE) Feed-forward Network in encoder blocks for multi-category adaptation.

– Comprehensive experimental results on the ShapeNet dataset demonstrate the state-of-the-art performance against the original DiT-3D while largely reducing the training costs.

2 Related Work

3D Shape Generation. The domain of 3D shape generation primarily revolves around creating high-quality point clouds through the utilization of generative models. These methods encompass various techniques, including variational autoencoders [12, 17, 32], generative adversarial networks [1, 27, 28], normalized flows [16,19,31], and Diffusion Transformers [23].

For example, Valsesia et al. [28] proposed a generative adversarial network leveraging graph convolution. Klokov et al. [19] introduced a latent variable model that employed normalizing flows to generate 3D point clouds. GET3D [13] used two latent codes to generate 3D signed distance functions (SDF) and textures, enabling the direct creation of textured 3D meshes.

Most recently, DiT-3D [23] pioneered the integration of denoising diffusion probabilistic models in the realm of 3D point cloud generation. Its efficacy in producing high-quality 3D point clouds has set a new benchmark in this domain, showcasing state-of-the-art performance. However, training voxel-based diffusion models for high-resolution 3D voxels ($128 \times 128 \times 128 \times 3$) remains prohibitively expensive due to the cubic complexity of attention operators, which arises from the additional dimension of voxels. Our focus is to explore methods for expediting the training process while upholding the generation quality. This exploration is critical to mitigate the computational constraints without compromising the fidelity of the generated outputs.

Diffusion Transformers in 3D Point Clouds Generation. Recent research, as documented in works such as [2,3,25,30], has highlighted the impressive performance of Diffusion Transformers. Diffusion Transformers have exhibited remarkable proficiency in generating high-fidelity images and even 3D point clouds, as outlined in [23]. In the area of image generation, the Diffusion Transformer (DiT) [25] presented a plain diffusion Transformer architecture aimed at learning the denoising diffusion process on latent patches. The U-ViT model [2] employed a Vision Transformer (ViT) [10]-based architecture with extensive skip connections.

In 3D point cloud generation, DiT-3D [23] presented a novel plain diffusion transformer tailored for 3D shape generation, specifically designed to perform denoising operations on voxelized point clouds effectively. This method achieved state-of-the-art performance and surpassed previous GAN-based or normalized flows-based methods by a large margin, demonstrating the effectiveness of diffusion transformer architecture in the 3D point cloud generation. However, it is worth noting that the training process is computationally expensive, prompting the exploration of methods to expedite and optimize the training phase.

Mask Diffusion Transformers. Transformers have emerged as predominant architectures in both natural language processing [9,29] and computer vision [11,

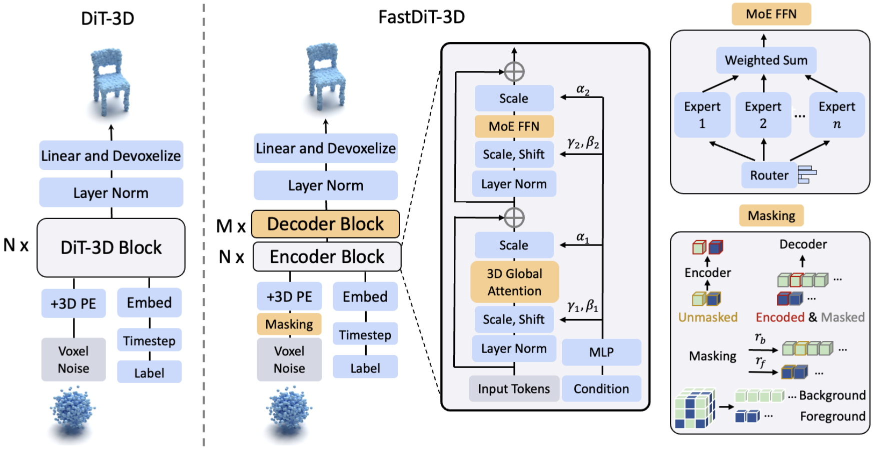

Fig. 2: Illustration of the proposed Fast training of Diffusion Transformers (FastDiT-3D) for 3D shape generation. The encoder blocks with 3D global attention and Mixture-of-Experts (MoE) FFN take masked voxelized point clouds as input. Then, multiple decoder transformer blocks based on 3D window attention extract point-voxel representations from all input tokens. Finally, the unpatchified voxel tensor output from a linear layer is devoxelized to predict the noise in the point cloud space.

[25]. The concept of masked training has found widespread application in generative modeling [5, 6, 26] and representation learning [9, 15, 20]. Within computer vision, a series of methodologies have adopted masked language modeling. MaskGiT [6] and MUSE [5] utilized the masked generative transformer for predicting randomly masked image tokens, enhancing image generation capabilities. MAE [15] further shows masked autoencoders are scaleable self-supervised learners. MDT [14] introduced a mask latent modeling scheme and achieved 3× faster learning speed than DiT [25]. MaskDiT [34] proposed an efficient approach to train large diffusion models with masked transformers by randomly masking out a high proportion of patches in diffused input images and achieves 31% of the training time of DiT [25]. Our work is the first to exploit masked training in the 3D point cloud generation domain. Even for a voxel size of $32 \times 32 \times 32$, our method achieves 10× faster training than the SOTA method DiT-3D [23] while exhibiting superior performance.

3 Method

Given a set of 3D point clouds, we aim to learn a plain diffusion transformer for synthesizing new high-fidelity point clouds. We propose a novel fast diffusion transformer that operates the denoising process of DDPM on masked voxelized point clouds, namely FastDiT-3D, which consists of two main modules: masked design DiT for 3D point cloud generation in Section 3.2 and Mixture-of-Experts encoder for multi-category generation in Section 3.3.

3.1 Preliminaries

In this section, we first describe the problem setup and notations and then revisit DDPMs for 3D shape generation and diffusion transformers on 3D point clouds.
Revisit DDPMs on 3D Shape Generation. In the realm of 3D shape generation, prior research, as exemplified by Zhou [23, 35], has leveraged DDPMs that involve a forward noising process and a reverse denoising process. In the forward pass, Gaussian noise is iteratively added to a real sample $\mathbf{x}_0$. By utilizing the reparameterization trick, $\mathbf{x}_t$ can be expressed as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$. $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, indicating the noise magnitude. If the timestep t is large, $\mathbf{x}_T$ would be a Gaussian noise. For the reverse process, diffusion models are trained to optimize a denoising network parameterized by $\boldsymbol{\theta}$ to map a Gaussian noise into a sample gradually. The training objective can be formulated as a loss between the predicted noise generated by the model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ and the ground truth Gaussian noise $\boldsymbol{\epsilon}$, denoted as $\mathcal{L}_{\text{simple}} = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)||^2$.

We train the diffusion model conditioned with class label, $p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t, c)$. During inference, new point clouds can be generated by sampling a Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then gradually denoise to obtain a sample $\mathbf{x}_0$.
Revisit DiT-3D on Point Clouds Generation. To address the generation challenge on inherently unordered point clouds, DiT-3D [23] proposed to voxelize the point clouds into dense representation in the diffusion transformers to extract point-voxel features. For each point cloud $\mathbf{p}_i \in \mathbb{R}^{N\times 3}$ with N points for x, y, z coordinates, DiT-3D first voxelized it as input $\mathbf{v}_i \in \mathbb{R}^{V\times V\times V\times 3}$, where V denotes the voxel size. Then, they applied the patchification operator with a patch size $p \times p \times p$ to generate a sequence of patch tokens $\mathbf{s} \in \mathbb{R}^{L\times 3}$, where $L = (V/p)^3$ is the total number of patchified tokens. Finally, several transformer blocks based on window attention were adopted to propagate point-voxel features. To achieve the denoising process in the point cloud space, the unpatchified voxel tensor is devoxelized into the output noise $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \in \mathbb{R}^{N\times 3}$.

Although DiT-3D [23] achieved promising results in generating high-fidelity 3D point clouds, they take the whole number L of patchified tokens as input to the encoder for feature propagation. The training process is computationally expensive, prompting the exploration of methods to expedite and optimize the training phase. Furthermore, the computational cost of 3D Transformers can be significantly high on the increased token length. Regarding high dimensions in 3D voxel space, such as 128 × 128 × 128, the training cost will be 1,668 A100 GPU hours. To address this challenge, we propose a novel fast plain diffusion transformer for 3D shape generation that can efficiently achieve the denoising processes on masked voxelized point clouds, as shown in Fig. 2.

3.2 DiT-3D for Masked Voxelized Point Clouds

Motivation. In order to achieve an efficient denoising process using a plain diffusion transformer during training, we propose several masked 3D design components in Figure 2 based on the SOTA architecture of DiT-3D [23] for 3D point cloud generation. Specifically, we introduce a novel foreground-background-aware

Category	Occupied	Non-occupied
Car	3.08%	96.91%
Chair	2.51%	97.49%
Airplane	1.42%	98.58%
Averaged	2.34%	97.66%

Table 1: Ratio Statistics on occupied (foreground) and non-occupied (background) voxels for different categories. A significant ratio gap between foreground and background voxels exists.

masking mechanism designed to mask voxelized point clouds as input. Such a novel strategy makes the masking ratio extremely high at nearly 99%, effectively leveraging the high inherent redundancy present in 3D data. We also replace 3D window attention with 3D global self-attention in the encoder blocks to propagate point-voxel representations from all unmasked tokens and add multiple decoder blocks with 3D window attention to take all patches tokens to predict the noise in the point cloud space. Finally, we apply a denoising objective on unmasked patches and a masked point cloud objective on masked patches for training our fast diffusion transformer on 3D point cloud generation.

Voxelized Point Clouds Masking. For a voxel of resolution $V \times V \times V$ with a total length of $L = (V/p)^3$, we apply a foreground-background masking mechanism to selectively filter out a substantial portion of patches, allowing only the remaining unmasked patches to proceed to the diffusion transformer encoder. Our observations reveal a significant ratio disparity between occupied and non-occupied voxels, as depicted in Table 1. Considering that occupied voxels contain information richness while background voxels are information-poor, we propose treating voxels in the occupied and non-occupied regions differently to optimize the masking ratio and attain the highest training efficiency. Specifically, we apply a ratio of r_f and a ratio of r_b to mask foreground patches $\mathbf{s}_f \in \mathbb{R}^{L_f \times 3}$ in occupied voxels and background patches $\mathbf{s}_b \in \mathbb{R}^{L_b \times 3}$ in non-occupied voxels, respectively. Therefore, we only pass $L_u = L - \lfloor r_f L_f \rfloor - \lfloor r_b L_b \rfloor$ unmasked patches to the diffusion transformer encoder. Our masking approach differs from random masking in image-based diffusion transformers [34]. Meanwhile, we empirically observe that the direct extension of MaskDiT [34] on point clouds does not work well, as random masking cannot select meaningful voxels for feature aggregation during the denoising process. Benefit from the masking strategy, our method is remarkably efficient that an extreme masking ratio r_b (*i.e.*, 99%) of background patches could still achieve efficient denoising for diffusion steps because the non-occupied background is 97.66% of overall voxels of all three categories on average, as shown in Table 1.

Encoder Blocks with 3D Global Attention. For encoding point-voxel representations from all unmasked patches L_u, we apply multiple encoder blocks based on the global multi-head self-attention operators with each of the heads $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ having dimensions $L_u \times D$, where L_u is the length of input unmasked tokens. The global attention operator is formulated as: $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) =$

$\text{Softmax}(\frac{\mathbf{QK}^\top}{\sqrt{D_h}}\mathbf{V})$, where D_h denotes the dimension size of each head. With our extremely high masking ratio, L_u is 327, while L is 32,768 for $128 \times 128 \times 128$ input voxels. Thus, given $L_u \ll L$, the computational complexity will be largely reduced to $\mathcal{O}(L_u^2)$ for this encoding process compared to the original complexity $\mathcal{O}(L^2)$ for high voxel resolutions. The efficiency further improves when considering the use of higher-resolution voxel input.

Decoder Blocks with 3D Window Attention. During the decoding process, we need to take all encoded unmasked tokens and masked tokens together, which leads to highly expensive complexity $\mathcal{O}(L^2)$ on the increased token length in 3D space. The computational cost of 3D Transformers can be significantly high. To alleviate this challenge, we are inspired by the original DiT-3D [23] and introduce efficient 3D window attention into decoder blocks to propagate point-voxel representations for all input patch tokens using efficient memory.

Specifically, we use a window size R to reduce the length of total input tokens $\hat{P}$ as follows. We first reshape $\hat{P}$ as: $\hat{P} : L \times D \rightarrow \frac{L}{R^3} \times (D \times R^3)$. And then apply a linear layer $\text{Linear}(C_{in}, C_{out})(\cdot)$ to $\hat{P} : P = \text{Linear}(D \times R^3, D)(\hat{P})$. And P denotes the reduced input patch tokens with a shape of $\frac{L}{R^3} \times D$. Therefore, the complexity of this decoding process is reduced from $\mathcal{O}(L^2)$ to $\mathcal{O}(\frac{L^2}{R^3})$.

Training Objectives. To achieve efficient training using our FastDiT-3D for masked 3D point clouds, we apply a denoising objective $\mathcal{L}_{\text{denoising}}$ on unmasked patches to use a mean-squared loss between the decoder output $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ and the ground truth Gaussian noise $\boldsymbol{\epsilon}$, and the objective is simply defined as $\mathcal{L}_{\text{denoising}} = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\|^2$. To make the model understand the global shape, we also utilize a masked point cloud objective $\mathcal{L}_{\text{mask}}$ on masked patches to minimize the mean-squared loss between the decoder output $\hat{\boldsymbol{\epsilon}}$ and the ground truth Gaussian noise $\boldsymbol{\epsilon}$ at current step t for masked patches. $\mathcal{L}_{\text{mask}} = \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\|^2$. Suppose a foreground-background aware mask $\boldsymbol{m} \in \{0, 1\}^L$, the overall objective is formulated as,

$$\begin{aligned} \mathcal{L} =& E_t(\|(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) \odot (1 - \boldsymbol{m})\|^2 + \\ & \lambda \cdot \|(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}) \odot \boldsymbol{m}\|^2) \end{aligned} \tag{1}$$

where $E_t(\|...\|^2 + \|...\|^2)$ represents the loss averaged across all timesteps, and λ denotes a coefficient to balance the denoising objective and masked prediction. In our experiments, we set it to 0.1 in default. Optimizing the denoising and masked loss together will push the learned representations of our FastDiT-3D to capture global 3D shapes for point cloud generation.

3.3 Mixture-of-Experts for Multi-class Generation

When trained on multi-category point clouds using one single dense model, the generation results will degrade compared to separately trained class-specific models. To improve the capacity of multi-category 3D shape generation in a single model, we integrate the Mixture-of-Experts (MoE) design to make the dense model sparse. Specifically, we replace each encoder block's original Feed Forward Network (FFN) with a MoE FFN. Given a router network $\mathcal{R}$ and several experts,

Method	Chair				Airplane				Car			
	1-NNA (↓)		COV (↑)		1-NNA (↓)		COV (↑)		1-NNA (↓)		COV (↑)	
	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD
r-GAN [1]	83.69	99.70	24.27	15.13	98.40	96.79	30.12	14.32	94.46	99.01	19.03	6.539
l-GAN (CD) [1]	68.58	83.84	41.99	29.31	87.30	93.95	38.52	21.23	66.49	88.78	38.92	23.58
l-GAN (EMD) [1]	71.90	64.65	38.07	44.86	89.49	76.91	38.27	38.52	71.16	66.19	37.78	45.17
PointFlow [31]	62.84	60.57	42.90	50.00	75.68	70.74	47.90	46.41	58.10	56.25	46.88	50.00
SoftFlow [16]	59.21	60.05	41.39	47.43	76.05	65.80	46.91	47.90	64.77	60.09	42.90	44.60
SetVAE [17]	58.84	60.57	46.83	44.26	76.54	67.65	43.70	48.40	59.94	59.94	49.15	46.59
DPF-Net [19]	62.00	58.53	44.71	48.79	75.18	65.55	46.17	48.89	62.35	54.48	45.74	49.43
DPM [22]	60.05	74.77	44.86	35.50	76.42	86.91	48.64	33.83	68.89	79.97	44.03	34.94
PVD [35]	57.09	60.87	36.68	49.24	73.82	64.81	48.88	52.09	54.55	53.83	41.19	50.56
LION [33]	53.70	52.34	48.94	52.11	67.41	61.23	47.16	49.63	53.41	51.14	50.00	56.53
GET3D [13]	75.26	72.49	43.36	42.77	–	–	–	–	75.26	72.49	15.04	18.38
MeshDiffusion [21]	53.69	57.63	46.00	46.71	66.44	76.26	47.34	42.15	81.43	87.84	34.07	25.85
DiT-3D-XL [23]	**49.11**	50.73	52.45	54.32	62.35	58.67	53.16	54.39	48.24	49.35	50.00	56.38
FastDiT-3D-S (ours)	50.35 (+1.24)	**50.27** (-0.46)	**58.53** (+6.08)	**60.79** (+6.47)	**61.83** (-0.52)	**57.86** (-0.81)	**58.21** (+5.05)	**58.75** (+4.36)	**47.81** (-0.43)	**48.83** (-0.52)	**53.86** (+3.86)	**59.62** (+3.24)

Table 2: Comparison results (%) on shape metrics of our FastDiT-3D and state-of-the-art models. Our method significantly outperforms previous baselines in terms of all classes.

which formulated as multi-layer perceptions (MLP), $\mathcal{E}_1$, $\mathcal{E}_2$, ..., $\mathcal{E}_n$, where n is the number of experts. During encoding on the input representations $\mathbf{x}_t$ from different categories, the router $\mathcal{R}$ activates the top-k expert networks with the largest scores $\mathcal{R}(\mathbf{x}_t)_j$, where j denotes the expert index. In order to sparsely activate different experts, the number of selected experts k is fixed during training and much smaller than the total number of experts n. The expert distribution of our Mixture of Expert (MoE) FFN layers can be formulated as:

$$\begin{aligned} \mathcal{R}(\mathbf{x}_t) &= \mathrm{TopK}(\mathrm{Softmax}(g(\mathbf{x}_t)), k) \\ \text{MoE-FFN}(\mathbf{x}_t) &= \sum_{j=1}^{k} \mathcal{R}(\mathbf{x}_t)_j \cdot \mathcal{E}_j(\mathbf{x}_t) \end{aligned} \tag{2}$$

where $\mathcal{E}_j(\mathbf{x}_t)$ denotes the representations from the expert $\mathcal{E}_j$, and $g(\cdot)$ is a learnable MLP within the router $\mathcal{R}$. TopK denotes an operator to select the top k ranked elements with the largest scores from $g(\cdot)$. By optimizing these experts to balance different categories during training, our FastDiT-3D further achieves adaptive per-sample specialization to generate high-fidelity 3D point clouds for multiple categories. Each class in this design is capable of capturing a unique diffusion path, involving a variety of experts across various layers. This approach significantly eases the challenge of complex gradient optimization that often arises from multi-class joint training.

3.4 Relationship to MaskDiT [34]

Our FastDiT-3D contains multiple different and efficient designs for 3D shape generation compared with MaskDiT [34] on 2D image generation:

- We utilize a foreground-background aware masking mechanism with an extremely high masking ratio of nearly 99%, while MaskDiT [34] adopted random masking with a relatively low masking ratio of 50%.
- Our FastDiT-3D performs efficient denoising on voxelized point clouds, while MaskDiT [34] needs the latent codes from a pre-trained variational autoencoder as the masked denoising target.

- We are the first to propose an encoder-decoder diffusion transformer on masked 3D voxelized point clouds for generating high-fidelity point clouds.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior works [23, 33, 35], we used ShapeNet [4] datasets, specifically focusing on the categories of Chair, Airplane, and Car, to serve as our primary datasets for the task of 3D shape generation. For a fair comparison with previous methods, we sampled 2,048 points from the 5,000 points provided within the ShapeNet dataset [4] for training and testing. For a fair comparison with previous approaches [23,33,35] on 3D shape generation, we follow the same procedures as outlined in PointFlow [31] for data preprocessing, which entails global data normalization applied uniformly across the entire dataset.
Evaluation Metrics. For comprehensive comparisons, we adopted the same evaluation metrics called Chamfer Distance (CD) and Earth Mover's Distance (EMD), as in prior methods [23,33,35], These metrics are instrumental in computing two key performance indicators: 1-Nearest Neighbor Accuracy (1-NNA) and Coverage (COV), which serve as primary measures of generative quality. 1-NNA computes the leave-one-out accuracy of the 1-Nearest Neighbor (1-NN) classifier to evaluate point cloud generation performance. This metric offers robust insights into the quality and diversity of generated point clouds, with a lower 1-NNA score indicating superior performance. COV quantifies the extent to which generated shapes match reference point clouds, serving as a measure of generation diversity. While a higher COV score is desirable, it's important to note that COV primarily reflects diversity and doesn't directly measure the quality of the generated point clouds. Therefore, it's possible for low-quality but diverse generated point clouds to achieve high COV scores.
Implementation. Our implementation is based on the PyTorch [24] framework. The input voxel size is set to $32 \times 32 \times 32 \times 3$, where $V = 32$ represents the spatial resolution. We perform weight initialization in accordance with established practices, with the final linear layer initialized to zeros and other weights following standard techniques typically employed in Vision Transformers (ViT) [10]. The models are trained for a total of 10,000 epochs, utilizing the Adam optimizer [18] with a learning rate of $1e-4$. Additionally, we use a batch size of 128. In our experiments, we set the diffusion time steps to $T = 1000$. By default, we apply a small backbone architecture with a patch size of $p = 4$. Notably, global attention is incorporated into all encoder blocks, while 3D window attention is selectively applied to specific decoder blocks (*i.e.*, 1 and 3). The total number n of experts is 6 in our MoE experiments.

4.2 Comparison to State-of-the-art Works

In this work, we introduce a novel and highly effective diffusion transformer tailored for 3D shape generation. To assess the efficacy of our proposed DiT-3D,

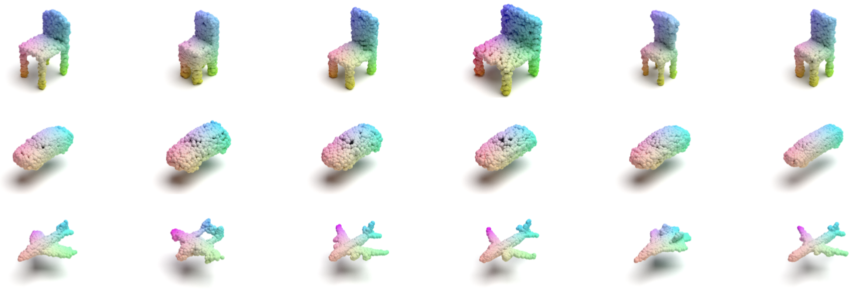

Fig. 3: Qualitative visualizations of high-fidelity and diverse 3D point cloud generation.

3D Voxel Masking	WA Decoder	Training Cost (hours)	1-NNA (↓) CD	EMD	COV (↑) CD	EMD
✗	✗	91	51.99	50.76	54.76	57.37
✓	✗	11	**50.09**	**50.02**	**59.79**	**61.45**
✓	✓	**8**	50.35	50.27	58.53	60.79

Table 3: Ablation studies on masked 3D components of our FastDiT-3D. Our model with both components has the lowest training costs while achieving competitive results.

we conduct a comprehensive comparative analysis against a range of baseline methods, encompassing both non-Diffusion Probabilistic Models (DDPM) [1,13,16,17,19,31], DDPM-based [21,22,33,35], and Diffusion Transformer-based [23] approaches.

We report the quantitative comparison results in Table 2. As can be seen, we achieved the best results regarding almost all metrics for both 1-NNA and COV evaluations compared to previous 3D shape generation approaches across the three categories. In particular, the proposed FastDiT-3D in model size of S remarkably superiorly outperforms DiT-3D [23] of model size XL, which is the current state-of-the-art diffusion transformer baseline.

Specifically, our method outperforms DiT-3D for airplane generation, decreasing by 0.52 in 1-NNA@CD and 0.81 in 1-NNA@EMD, and increasing by 5.05 in COV@CD and 4.36 in COV@EMD. Furthermore, we achieve significant performance gains compared to LION [33], a recent competitive baseline based on two hierarchical DDPMs. The results demonstrate the importance of masked prediction in capturing global 3D shapes for point cloud generation. In addition, significant gains in chair and car generation can be observed in Table 2. These significant improvements demonstrate the superiority of our approach in 3D point cloud generation. These qualitative results in Fig. 3 also showcase the effectiveness of the proposed FastDiT-3D in generating high-fidelity and diverse 3D point clouds.

r_b	r_f	Training	1-NNA (↓)		COV (↑)	
		Cost (hours)	CD	EMD	CD	EMD
Random masking:						
0%		91	51.99	50.76	54.76	57.37
50%		55	50.82	50.15	57.69	59.12
75%		31	51.32	50.46	58.03	59.37
95%		15	51.53	50.52	57.85	59.28
99%		11	82.35	85.16	29.63	23.56
Foreground-background aware masking:						
95%	95%	15	50.22	50.06	59.25	61.23
97%	95%	13	50.17	50.05	**59.86**	**61.53**
99%	95%	11	**50.09**	**50.02**	59.79	61.45
99%	**96%**	10.5	50.86	50.65	57.63	58.52
100%	95%	**10**	52.87	51.69	55.23	56.82

Table 4: Exploration studies on the trade-off of non-occupied (r_b) and occupied (r_f) masking ratio. When r_b, r_f is 99%, 95%, we achieve decent generation results and training costs together.

4.3 Experimental Analysis

In this section, we performed ablation studies to demonstrate the benefit of introducing two main 3D design components (3D voxel masking and 3D window attention decoder) in 3D shape generation. We also conducted extensive experiments to explore the efficiency of a mixture-of-experts encoder, modality domain transferability, and scalability.

Ablation on 3D Masked Design Components. In order to demonstrate the effectiveness of the introduced 3D voxel masking and 3D window attention (WA) decoder, we ablate the necessity of each module and report the quantitative results in Table 3. We can observe that adding 3D voxel masking to the vanilla baseline highly decreases the training hours from 91 to 11, and improves the generation results by reducing 1.90 in 1-NNA@CD and 0.74 in 1-NNA@EMD and increasing 5.03 in COV@CD and 4.08 in COV@EMD. Meanwhile, introducing the WA decoder further decreases the training hours to 8, while achieving competitive performance. These improving results validate the importance of 3D voxel masking and 3D window attention decoder on efficient training and effective 3D point cloud generation.

Trade-off of Non-occupied/occupied Masking Ratio. The number of non-occupied/occupied masking ratios used in the proposed 3D voxel masking module affects the extracted patch tokens for feature aggregation on point cloud generation. To explore such effects more comprehensively, we first varied the number of masking ratios from $\{0, 50\%, 75\%, 95\%, 99\%\}$ in random masking, and then ablated the non-occupied masking ratio r_b from $\{95\%, 97\%, 99\%, 100\%\}$ and occupied masking ratio r_f from $\{95\%, 96\%\}$. It should be noted that we do not discriminate non-occupied/occupied voxels for random masking, resulting in the same ratio for all voxels. The comparison results of chair generation are reported in Table 4. When the number of masking ratio is 99% for random masking, we

ImageNet Pre-train	Training Cost (hours)	1-NNA (↓) CD	1-NNA (↓) EMD	COV (↑) CD	COV (↑) EMD
✗	8	**50.35**	**50.27**	58.53	60.79
✓	7	50.39	50.28	**58.62**	**60.86**

(a) Modality transfer.

Mixture-of-experts	Params (MB)	1-NNA (↓) CD	1-NNA (↓) EMD	COV (↑) CD	COV (↑) EMD
✗	54.73	52.16	51.05	56.53	58.17
✓($k=1$)	58.26	51.95	50.87	56.86	58.63
✓($k=2$)	68.92	**51.72**	**50.56**	**57.38**	**59.26**

(b) Mixture-of-experts. Top k experts are selected.

Table 5: Ablation studies on 2D pretrain and Mixture-of-experts for multi-category generation.

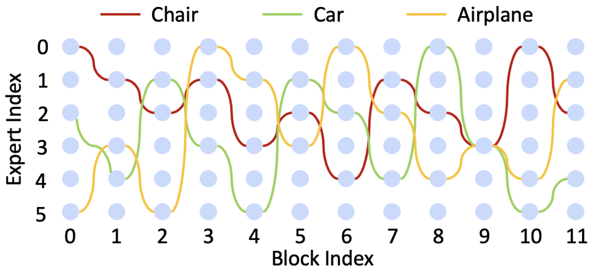

Fig. 4: Qualitative visualizations of sampling paths across experts in Mixture-of-Experts encoder blocks for multi-class generation. The learned various paths denote different classes. It demonstrates that each category can learn a distinct diffusion path.

achieve the lowest training costs but the model does not work. With the increase of non-occupied masking ratio r_b from 95% to 99%, the proposed FastDiT-3D consistently improves results in terms of generation quality. The superior performance on such an extreme masking ratio demonstrates the importance of foreground-background aware masking strategy which effectively optimizes the masking ratio and obtains the highest training efficiency. Moreover, we conduct experiments of increasing the non-occupied masking ratio r_b from 99% to 100% and increasing the occupied masking ratio r_b from 95% to 96%, the results will not continually improve. This is because there might be indispensable voxel patches in foreground and background for generating high-fidelity point clouds.

Influence of 2D Pretrain (ImageNet). 2D ImageNet pre-trained weights has been demonstrated effective in DiT-3D [23] for modality transferability to 3D generation with parameter-efficient fine-tuning. In order to explore such an effect of modality transferability on our FastDiT-3D, we initialized our encoder and decoder weights from MaskDiT [34] and continued to fine-tune all parameters during training. The ablation results on chair generation are reported in Table 5a. We can observe that using ImageNet pre-trained weights achieves fast convergence with fewer training hours and competitive results on high-fidelity point cloud generation, where it outperforms the original random initialization on COV metrics for generating diverse shapes.

Mixture-of-Experts FFN for Multi-class Generation. In order to demonstrate the effectiveness of mixture-of-experts FFN in our encoder blocks for generating high-fidelity point clouds from multiple categories, we varied the number of top selected experts k from $\{1, 2\}$, and report the comparison results in Table 5b. As can be seen, adding MoE FFN of one expert activated with similar parameters as our FastDiT-3D without MoE achieves better results in terms of all metrics. Increasing the number of activated experts further improves the performance but brings more training parameters. These improving results validate the importance of the mixture-of-experts FFN in generating high-fidelity point clouds. Fig. 4 also showcases the sample paths across different experts in MoE encoder blocks for multi-category generation for samples from chair, car, and airplane, where the index with the highest frequency of occurrence of experts in each layer are calculated on all training samples corresponding to each class. We can observe that each class is able to learn a distinct, unique diffusion path, which dynamically chooses different experts in different layers, improving the model's capacity to generate multiple categories.

5 Conclusion

In this work, we propose FastDiT-3D, a novel fast diffusion transformer tailored for efficient 3D point cloud generation. Compared to the previous DiT-3D approaches, Our FastDiT-3D dynamically operates the denoising process on masked voxelized point clouds, offering significant improvements in training cost of merely 6.5% of the original training cost. And FastDiT-3D achieves superior point cloud generation quality across multiple categories. Specifically, our FastDiT-3D introduces voxel-aware masking to adaptively aggregate background and foreground information from voxelized point clouds, thus achieving an extreme masking ratio of nearly 99%. Additionally, we incorporate 3D window attention into decoder Transformer blocks to mitigate the computational burden of self-attention in the context of increased 3D token length. We introduce Mixture of Expert (MoE) layers into encoder transformer blocks to enhance self-attention for multi-category 3D shape generation. Extensive experiments on the ShapeNet dataset demonstrate that the proposed FastDiT-3D achieves state-of-the-art generation results in high-fidelity and diverse 3D point clouds. We also conduct comprehensive ablation studies to validate the effectiveness of voxel-aware masking and 3D window attention decoder. Qualitative visualizations of distinct sampling paths from various experts across different layers showcase the efficiency of the MoE encoder in multi-category generation.

Limitations & Broader Impact. Although the proposed FastDiT-3D achieves superior results in generating high-fidelity and diverse point clouds given classes, we have not explored the potential usage of explicit text control for 3D shape generation. Furthermore, we can scale our FastDiT-3D to large-scale text-3D pairs [7,8] for efficient training on text-to-3D generation. These promising directions will leave for future work.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: Proceedings of the International Conference on Machine Learning (ICML) (2018) 4, 9, 11
2. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
3. Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., Zhu, J.: One transformer fits all distributions in multi-modal diffusion at scale. arXiv preprint arXiv:2303.06555 (2023) 4
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 10
5. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023) 2, 5
6. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022) 2, 5
7. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023) 14
8. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. arXiv preprint arXiv:2212.08051 (2022) 14
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4, 5
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations (ICLR) (2021) 4, 10
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
12. Gadelha, M., Wang, R., Maji, S.: Multiresolution tree networks for 3d point cloud processing. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 4
13. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. In: Proceedings of Advances In Neural Information Processing Systems (NeurIPS) (2022) 4, 9, 11
14. Gao, S., Zhou, P., Cheng, M.M., Yan, S.: Masked diffusion transformer is a strong image synthesizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23164–23173 (2023) 5

15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022) 2, 3, 5
16. Kim, H., Lee, H., Kang, W., Lee, J.Y., Kim, N.S.: Softflow: Probabilistic framework for normalizing flow on manifolds. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020) 4, 9, 11
17. Kim, J., Yoo, J., Lee, J., Hong, S.: Setvae: Learning hierarchical composition for generative modeling of set-structured data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15059–15068 (2021) 4, 9, 11
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
19. Klokov, R., Boyer, E., Verbeek, J.: Discrete point flow networks for efficient point cloud generation. In: Proceedings of the European Conference on Computer Vision (ECCV). p. 694–710 (2020) 4, 9, 11
20. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pretraining via masking. arXiv preprint arXiv:2212.00794 (2022) 5
21. Liu, Z., Feng, Y., Black, M.J., Nowrouzezahrai, D., Paull, L., Liu, W.: Meshdiffusion: Score-based generative 3d mesh modeling. In: Proceedings of International Conference on Learning Representations (ICLR) (2023) 2, 9, 11
22. 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 (CVPR). pp. 2837–2845 (2021) 9, 11
23. Mo, S., Xie, E., Chu, R., Yao, L., Hong, L., Nießner, M., Li, Z.: DiT-3D: Exploring plain diffusion transformers for 3d shape generation. In: Proceedings of Advances In Neural Information Processing Systems (NeurIPS) (2023) 1, 4, 5, 6, 8, 9, 10, 11, 13
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). pp. 8026–8037 (2019) 10
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022) 4, 5
26. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. OpenAI (2018) 5
27. Shu, D.W., Park, S.W., Kwon, J.: 3d point cloud generative adversarial network based on tree structured graph convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3859–3868 (2019) 4
28. Valsesia, D., Fracastoro, G., Magli, E.: Learning localized generative models for 3d point clouds via graph convolution. In: Proceedings of International Conference on Learning Representations (ICLR) (2019) 4
29. 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) 4
30. Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J., Li, Z.: Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. arXiv preprint arXiv:2304.06648 (2023) 4

31. 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 (ICCV). pp. 4541–4550 (2019) 4, 9, 10, 11
32. Yang, Y., Feng, C., Shen, Y., Tian, D.: Foldingnet: Point cloud auto-encoder via deep grid deformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 206–215 (2018) 4
33. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022) 1, 9, 10, 11
34. Zheng, H., Nie, W., Vahdat, A., Anandkumar, A.: Fast training of diffusion models with masked transformers (2023) 2, 3, 5, 7, 9, 13
35. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5826–5835 (2021) 6, 9, 10, 11