Generalized Deep 3D Shape Prior via Part-Discretized Diffusion Process

Yuhan Li[1] Yishun Dou[2] Xuanhong Chen[1] Bingbing Ni[1, 2†] Yilin Sun[1] Yutian Liu[1] Fuzhen Wang[1]
[1]Shanghai Jiao Tong University, Shanghai 200240, China [2]Huawei
{melodious, nibingbing}@sjtu.edu.cn

https://github.com/colorful-liyu/3DQD

Abstract

We develop a generalized 3D shape generation prior model, tailored for multiple 3D tasks including unconditional shape generation, point cloud completion, and cross-modality shape generation, etc. *On one hand, to precisely capture local fine detailed shape information, a vector quantized variational autoencoder (VQ-VAE) is utilized to index local geometry from a compactly learned codebook based on a broad set of task training data. On the other hand, a discrete diffusion generator is introduced to model the inherent structural dependencies among different tokens. In the meantime, a multi-frequency fusion module (MFM) is developed to suppress high-frequency shape feature fluctuations, guided by multi-frequency contextual information. The above designs jointly equip our proposed 3D shape prior model with high-fidelity, diverse features as well as the capability of cross-modality alignment, and extensive experiments have demonstrated superior performances on various 3D shape generation tasks.*

1. Introduction

While pre-trained 2D prior models [24, 43] have shown great power in various downstream vision tasks such as image classification, editing and cross-modality generation, *etc*., their counterpart 3D prior models which are generally beneficial for three-dimensional shape generation tasks have NOT been well developed, unfortunately. On the contrary, the graphics community has developed a number of task-specific pre-trained models, tailored for unary tasks such as 3D shape generation [22, 31, 60], points cloud completion [62, 66, 67] and conditional shape prediction [14, 30, 34, 54]. Since above individual 3D generative representations do NOT share common knowledge among tasks, to migrate a trained 3D shape network from one task to another related one requires troublesome end-to-end model re-work and training resources are also wasted. For instance, a good shape encoding of "chairs" based on a general prior 3D model could benefit shape completion of a given partial chair and text-guided novel chair generation.

†Corresponding author: Bingbing Ni.

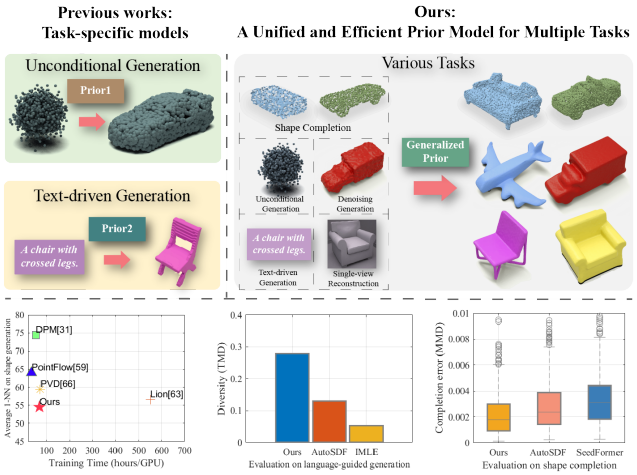

Figure 1. Our shape generation model is a unified and efficient prior model to produce high-fidelity, diverse results on multiple tasks, while most previous approaches are task-specific.

This work aims at designing a unified 3D shape generative prior model, which serves as a generalized backbone for multiple downstream tasks, *i.e*., with few requirements for painstaking adaptation of the prior model itself. Such a general 3D prior model should possess the following *good* properties. On one hand, it should be *expressive* enough to generate **high-fidelity** shape generation results with fine-grained local details. On the other hand, this general prior should cover a large probabilistic support region so that it could sample **diverse** shape prototypes in both conditional and unconditional generation tasks. Moreover, to well support **cross-modal** generation, *e.g*., text-to-shape, the sampled representation from the prior model should achieve good semantic consistency between different modalities, making it easy to encode and match partial shapes, images and text prompts.

The above criteria, however, are rarely satisfied by existing 3D shape modeling approaches. Encoder-decoder based structures [62, 66] usually focus on dedicated tasks and fail to capture diverse shape samples because the generation process mostly relies on features sampled from deterministic encoders. On the contrary, probabilistic models such as GAN [29, 48], Flow [60] and diffusion models [22, 31, 67],

cannot either adapt to multiple tasks flexibly or generate high-quality shapes without artifacts [64]. Note that recent models such as AutoSDF [34] and Lion [64] explore multiple conditional generative frameworks. However, AutoSDF [34] is defective in diversity and shows mode collapse in unconditional and text-guided generation, while training and inference of Lion [64] are costly.

In view of above limitations, this work proposes an efficient ***3D-Disassemble-Quantization-and-Diffusion (3DQD)*** prior model to simultaneously address above challenges in 3D shape generation: high-fidelity, diversity, and good cross-modality alignment ability, as shown in Fig. 1. This 3DQD prior is a unified, probabilistic and powerful backbone for both unconditional shape generation and multiple conditional shape completion applications. On one hand, instead of using holistic codes for whole-shape, we adopt a vector quantized variational autoencoder [52] (VQ-VAE) that learns a compact representation for *disassembled* local parts of the shape, which is expected to well represent diverse types of local geometry information out of a broad set of training shapes from shared tasks, forming a generalized part-level codebook. Note that this disassembled and quantized representation effectively eliminates the intrinsic structural bias between different modalities, and thus is able to perform good cross-modality data alignment on conditional generation tasks. On the other hand, a discrete diffusion generator [4, 16] with reverse Markov chain is introduced to model the inherent semantic dependencies among geometric tokens. Namely, the forward process corrupts the latent variables with progressively increasing noise, transferring parts of variables into random or masked ones; and the reverse process gradually recovers the variables towards the desired data distribution by learning the structural connections among geometry tokens. It is worth mentioning that random corruption in the forward process facilitates diverse samples, while discrete embedding results in a dramatic cost-down in computational budget, with stable and iterative sampling. Furthermore, during shape generation we introduce **Multi-frequency Fusion Module** (MFM) to suppress high-frequency outliers, encouraging smoother and high-fidelity samples.

Comprehensive and various downstream experiments on shape generation tasks demonstrate that the proposed 3DQD prior is capable of helping synthesize high-fidelity, diverse shapes efficiently, outperforming multiple state-of-the-art methods. Furthermore, our 3DQD prior model can serve as a generalized backbone with highly competitive samples for extended applications and cross-domain tasks requiring NO or little tuning.

2. Related Works

Diffusion models. Recently, the denoising diffusion probabilistic model (DDPM), has attracted great attention in the community [18, 35, 43], which utilizes a Markov chain to convert the noise distribution to the data distribution. Models under this family have achieved strong results on image generation [3, 12, 18, 19, 36, 44] and multiple conditional generation tasks [9, 11, 33, 37, 43, 45, 46]. Meanwhile, other researchers investigated the discrete diffusion models. Argmax Flow [21] first introduces a diffusion model directly defined on discrete categorical variables, while D3PM [4] generalizes it by going beyond corruption processes with uniform transition probabilities. VQ-Diffusion [16,50] encodes images into discrete feature maps to train the diffusion, presenting comparable results in text-to-image synthesis with the state-of-the-art models [35,41].

3D Shape generation. Traditional shape generation methods [15, 29, 61] mitigate generation process by mapping a primitive matrix to a point cloud with transformation. Some recent works [5, 31, 60, 67] consider point clouds as samples from a distribution by different probabilistic models. Almost all of them cannot cope with conditional input and multi-modality data. Some conditional works [10, 51, 62, 63, 65, 66] focus on completing full shapes from partial inputs on point clouds with task-specific architectures to encode inputs and decode the global features. Others explore single-view reconstruction [32, 47, 54–57] and text-driven generation [8, 14, 30, 34] to learn a joint condition-shape distribution with deterministic process. However, most conditional generation methods fail in capturing the multiple output modes [34] and adapting to various tasks.

Most related to our work, PVD [67] and Lion [64] both use a diffusion model for diverse generation. Their models work on latent embedding space, leading to either difficulty in multi-modality applications (*i.e.*, denosing) [64] or time-consuming training and inference. Inspired by discrete approaches [13], we introduce VQ-VAE [52] to learn a compact representation, followed by a discrete diffusion generator, to reduce computational overhead and align cross-domain data. Our model is a superior and efficient prior model for shape generation compared with PVD [67] and Lion [64], and shows more diversity than AutoSDF [34].

3. Methodology

In this section, we introduce in detail of our proposed general 3D shape prior model (3DQD) for various 3D generation tasks. The architecture is visualized in Fig. 2. To begin with, we present our shape encoding scheme based on part-level discretization using VQ-VAE [52] style methods, for its advantages in representation compactness and consistency among different tasks. Then, a novel diffusion process based prior is developed according to this discrete part based shape representation with expressive and diversified object structural modeling, forming a good basis for 3D generation downstream applications. Moreover, a novel multi-frequency fusion module is introduced for enhancing fine-grained 3D surface modeling.

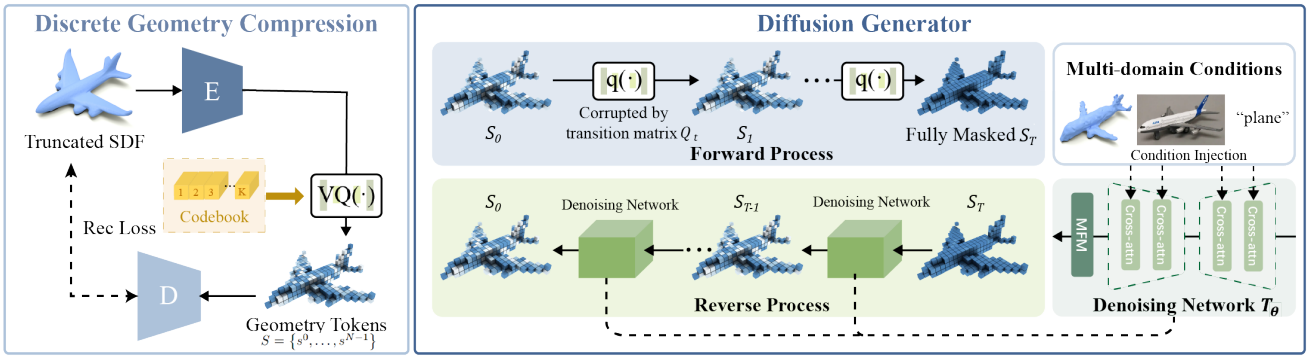

Figure 2. Overall framework of our method. VQ-VAE encodes shapes into geometry tokens with compact representation. Then diffusion generator models the joint distribution of tokens by reversing a forward diffusion process that corrupts the inputs via a Markov chain.

3.1. Discrete Part-based Geometry Encoding

A proper representation (*i.e.*, shape encoding) is essential for effective shape distribution model design. Previous coding schemes [31, 67] based on continuous embedding representation learning suffer from drawbacks of large computational consumption and heavy task dependence, rendering difficulty in 3D conditional shape generation, especially for multi-task settings. While most encoder-based methods (*i.e.* Lion [64]) choose to embed shapes into continuous space via VAE [27], in this work, we propose to circumvent this drawback by introducing Patch-level VQ-VAE (P-VQ-VAE) [34, 52] on Truncated-Signed Distance Field [23, 58] (T-SDF) to learn a compact and efficient representation. The motivation is obvious. All 3D objects are composed of geometrically similar local components, and thus a good local atomic representation can be shared among different objects and diverse tasks.Besides, using discrete codebook to index each local part results in a dramatic cost-down in searching discrete feature space. As it's flexible to encode inter-dependencies between parts, this encoding scheme provides a compact way for expressing highly diverse intrinsic structural variations. Moreover, local atomic indexing potentially offers unified and well-aligned local representation among different modalities, *e.g.* images and text prompts, which greatly facilitate cross-domain conditional shape generation.

Formally, the proposed P-VQ-VAE consists of an encoder E, a decoder D and codebook entries $\mathcal{Z} \in \mathbb{R}^{K \times n_z}$, which contains a finite number of embedding vectors, where K is the size of codebook and n_z is the dimension of vectors. Following AutoSDF [34], given a T-SDF input $X \in \mathbb{R}^{H \times W \times D}$, the whole 3D shape is divided into partial regions $X' = \{x'^0, \ldots, x'^{N-1}\}$, where $x'^i \in \mathbb{R}^{h \times w \times d}$ and N is the number of regions. Afterwards, each part is vectorized by the encoder as $z^i = E(x'^i) \in \mathbb{R}^{n_z}$. Superscript i is its spatial location. Finally, we obtain a spatial collection of quantized shape tokens $Z_q = \{z_q^0, \ldots, z_q^{N-1}\}$ by a further quantization step as $z_q^i = VQ(z^i) \in \mathbb{R}^{n_z}$, and their corresponding codebook indexes $S = \{s^0, \ldots, s^{N-1}\}$, where $s^i \in \{0, \ldots, K-1\}$. VQ is a spatial-wise quantizer which maps partial shape z^i into its closest codebook entry in $\mathcal{Z}$.

3.2. 3D Shape Prior: Discrete Diffusion Model

Given the above P-VQ-VAE encoding (*i.e.*, discrete part geometry tokens $Z_q \in \mathbb{R}^{N \times n_z}$ and corresponding indexes $S = \{s^0, \ldots, s^{N-1}\}$), to build the 3D shape prior is equivalent to modeling the **joint probability distribution** of all local shape codes in the latent space. In essence, a sampling from this prior distribution reveals a certain intrinsic structure of a 3D shape, by considering the inter-relationship geometric organization among local shape codes, *i.e.*, how to spatially combine local parts into various 3D objects. In this sense, a good prior model, should not only provide a probabilistic support region as wide as possible (*i.e.* possessing sufficient shape diversity), but also be general enough to deal with different downstream 3D shape tasks. For example, a chair with 4 legs can be transformed into a chair with wheels easily by the knowledge learned in generation, without training on target editing tasks.

The joint distribution of partial components and conditions are usually learned with autoregressive models in previous works, *e.g.*, DALL-E [42], Taming [13], ShapeFormer [59] and AutoSDF [34], along with sequential non-Markov-chain styled generation process: $\prod_{i=1}^{N} p_\theta(s^i | s^1, \ldots, s^{i-1})$. However, this scheme has several drawbacks. 1) **Error Accumulation**: Geometry tokens are predicted one by one, therefore errors induced in the earlier sampling timesteps will never be corrected and contaminate the subsequent generation steps; 2) **Un-optimal Generation Order**: Most autoregressive models perform an unidirectional prediction process, *e.g.* left-to-right, bottom-to-top, front-to-back or random orders, which obviously ignores the underlying complex 3D structure; and 3) **Lack of Diversity**: Deterministic transformers are usually instantiated as backbones of autoregressive models [34, 59], and without sufficient randomness injection they easily lean towards mode collapse (*e.g.*, highly similar completion results

given a partial shape), especially in condition-driven tasks.

In view of these limitations, we develop a discrete diffusion generator [16, 50] to iteratively sample in the time domain with all local part codes updated simultaneously at each timestep (*i.e.*, instead of sampling each spatial location one by one). In this way, diffusion generator is able to get rid of fixed unidirectional generation and update all partial geometry with long-range dependencies simultaneously, which enhances structural expressivity of the learned shape distribution. Within this framework, earlier samples are rechecked multiple times during iterations, reducing the likelihood of being confused. In addition, random corruption on shapes in diffusing process also leads to a great diversity of generated results. The forward sampling process for shape generation and the backward sampling process for training the proposed discrete diffusion model are introduced in detail as follows.

3.2.1 Forward Diffusing Process and Corruption.

The forward diffusion process gradually adds noise to tokenized 3D geometric input $S_0 = \{s_0^0, \dots, s_0^{N-1}\}$ (superscripts are locations and subscripts are timesteps) via a Markov chain $q(s_{1:T}^i|s_0^i) = \prod_{t=1}^{T} q(s_t^i|s_{t-1}^i)$, where each token is randomly replaced into noisy index $s_{1:T}^i = s_1^i, s_2^i, \dots, s_T^i$. Without introducing confusion, we omit superscripts i in the following description. After T timesteps, each token in the entire map will be completely corrupted into s_T, *i.e.*, a non-sense index. The learned reverse process $p_\theta(s_{0:T}) = p(s_T)\prod_{t=1}^{T} p_\theta(s_{t-1}|s_t)$ gradually removes the added noise on the random variables s_t, namely to generate a 3D shape from random noise.

While continuous diffusion process models [18, 64, 67] employ Gaussian noise in the forward process; in discrete diffusion process, we use transition matrices $[Q_t]_{i,j} = q(s_t = j|s_{t-1} = i)$ characterized by uniform transition probabilities to describe the corruption from s_{t-1} to s_t. As a result, all local part codes can be transformed to any other shape codes with the same transition probability, producing complex 3D structure representations. With transition matrix $[Q_t]$ and one-hot encoding of s_t, we define the forward Markov chain as:

$$q(s_t|s_{t-1}) = \mathbf{\Psi}(s_t; p = s_{t-1}Q_t), \tag{1}$$

where $\mathbf{\Psi}(s_t; p)$ is a categorical distribution over the one-hot row vector s_t sampled with probability p, and $s_{t-1}Q_t$ is computed by a row vector-matrix product. Accordingly, we derive the posterior by Markov chain iteratively from s_0 as:

$$q(s_t|s_0) = \mathbf{\Psi}(s_t; p = s_0\overline{Q_t}), \text{ with } \overline{Q_t} = Q_1 \cdots Q_t, \tag{2}$$

$$\begin{aligned} q(s_{t-1}|s_t, s_0) &= \frac{q(s_t|s_{t-1}, s_0)q(s_{t-1}|s_0)}{q(s_t|s_0)} \\ &= \mathbf{\Psi}\left(s_{t-1}; p = \frac{s_t Q_t^\top \odot s_0 \overline{Q}_{t-1}}{s_0 \overline{Q}_t s_t^\top}\right). \end{aligned} \tag{3}$$

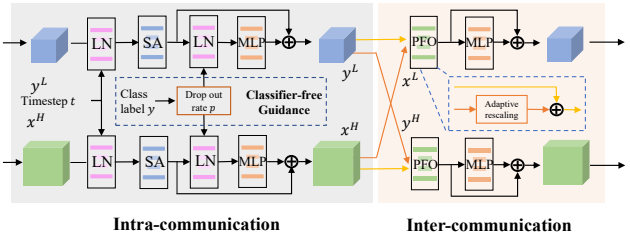

Figure 3. Information from detailed shapes and condensed features are intra- and inter-communicated in MFM to remove high-frequency components from uniform noise, where timestep t and classifier-free guidance are modulated with AdaLayerNorm.

The transition matrix Q_t reflects how to spatially reorganize 3D local primitives and it determines the degree of freedom in deformation. In order to assist the network to locate what needs to be fixed quickly, we introduce an additional special token, [*MASK*] token [16]. So the codebook entries are composed of $K+1$ states: K geometry tokens and one [*MASK*] token. $Q_t \in \mathbb{R}^{(K+1)\times(K+1)}$ can be formulated as:

$$[Q_t] = \begin{bmatrix} \alpha_t - \frac{K-1}{K}\beta_t & \frac{\beta_t}{K} & \frac{\beta_t}{K} & \cdots & \gamma_t \\ \frac{\beta_t}{K} & \alpha_t - \frac{K-1}{K}\beta_t & \frac{\beta_t}{K} & \cdots & \gamma_t \\ \frac{\beta_t}{K} & \frac{\beta_t}{K} & \alpha_t - \frac{K-1}{K}\beta_t & \cdots & \gamma_t \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{4}$$

with $\alpha_t = 1 - \gamma_t$, which means each token has a probability of $1 - \gamma_t - \frac{K-1}{K}\beta_t$ to remain unchanged, while with a probability of $\frac{\beta_t}{K}$ being transited equally into other $K-1$ geometry categories and γ_t into [*MASK*] state.

3.2.2 Reverse Sampling Process and Model Learning

To recover original 3D grids of tokens from s_t, we follow Ho *et al*. [18] and Hoogeboom *et al*. [21], and train a denoising transformer T_θ to directly estimate the noise-removed shape distribution $p_\theta(\hat{s}_0|s_t)$. Then the posterior transition distribution can be obtained with $q(s_{t-1}|s_t, s_0)$ as:

$$p_\theta(s_{t-1}|s_t) = \sum_{\hat{s}_0} q(s_{t-1}|s_t, \hat{s}_0)p_\theta(\hat{s}_0|s_t). \tag{5}$$

The evidence lower bound (ELBO) associated with this model then decomposes over the discrete timesteps as:

$$\begin{aligned} -\log p_\theta(\hat{s_0}) \leq\ & \mathbb{KL}(q(s_T|s_0)|p(s_T)) \\ & + \sum_{t=1}^{T} \mathbb{E}_{q(s_t|s_0)}\mathbb{KL}(q(s_{t-1}|s_t, s_0)|p_\theta(s_{t-1}|s_t)). \end{aligned} \tag{6}$$

For the first term of ELBO has no relation with the learnable network ($p(s_T)$ is an initial noise distribution), the common way to minimize ELBO is to optimize the posterior $p_\theta(s_{t-1}|s_t)$ in terms of Eq. (5). Due to the direct prediction of neural network is original shapes $p_\theta(\hat{s}_0|s_t)$, we set an auxiliary training objective with the distance of recovered original shapes apart from the posterior, yielding the following loss function:

$$L = -\mathbb{E}_{q(s_0)q(s_t|s_0)} \left[\underbrace{\log p_\theta(s_{t-1}|s_t)}_{L_{main}} + \lambda \underbrace{\log p_\theta(\hat{s}_0|s_t)}_{L_{aux}} \right], \quad (7)$$

where λ balances the weight of the auxiliary loss.

Classifier-free Guidance. To trade off between precision and diversity, Ho & Nichol *et al*. [20, 35] propose classifier-free guidance, which does not require a separately trained classifier model. During training, the class label y in diffusion model has a fixed probability (drop out rate p) to be replaced with the empty label $\emptyset$. During inference, the model prediction is thus adjusted in the direction of $p_\theta(\hat{s}_0|s_t, y)$ away from $p_\theta(\hat{s}_0|s_t, \emptyset)$ as:

$$p_\theta(\hat{s}_0|s_t) = (1+w) \cdot p_\theta(\hat{s}_0|s_t, y) - w \cdot p_\theta(\hat{s}_0|s_t, \emptyset), \quad (8)$$

with $w \geq 0$ controlling the guidance scale.

In this work, we also employ classifier-free guidance on multi-category generation tasks (*e.g*., shape completion), which is proved to enrich diversity of samples without apparently affecting the fidelity. We set drop out rate $p = 0.5$ and guidance weight $w = 0.5$.

3.2.3 Multi-frequency Fusion Module

In the meantime, we observe noisy surfaces with outliers from generated shapes mainly introduced by *categorical corruption* in discrete diffusion models. Namely, our *categorical corruption* has equal probability to transit a token into any completely irrelevant shape tokens [4], and it is hard to recover the correct category of this shape token without looking into its contextual information, *e.g*., its adjacent tokens. Note that in continuous diffusion models, the added Gaussian noise only brings up *soft* transition which still preserves part of the original shape information. Therefore, our transition matrix Q_t with uniform noise imports more high-frequency noise and outliers than Gaussian diffusion [18] and autoregressive model [13, 52].

To remedy this issue, inspired by the observation that corrupted tokens' neighbors are still possible to remain unchanged and provide dependencies, we develop a Multi-frequency Fusion Module (MFM) which looks into local contextual regions and extracts low-frequency components by downsampling to suppress high-frequency outliers, encouraging smoother and high-fidelity samples. As shown in Fig. 3, we split the detailed shape token embedding x^H explicitly with the condensed features y^L from its neighbors, which is obtained by down-sampling within its local receptive field. Each component is further sent into intra- and inter-relationship part: $[x^{H \longrightarrow H \longrightarrow L}]$ and $[y^{L \longrightarrow L \longrightarrow H}]$, respectively, and communication of information is realized. Specifically, we employ self-attention for intra-frequency update, and Pair-wise Fusion Operator (PFO) for communication between two components. We examine two ways of communication (*i.e*. cross-attention [7], residual add [17]). Note that residual add

$$x^H + f(x^H + \mathcal{P}(y^L)), \quad (9)$$

Method	Airplane		Chair		Car	
	CD↓	EMD↓	CD↓	EMD↓	CD↓	EMD↓
r-GAN [1]	99.84	96.79	83.69	99.70	94.46	99.01
PointFlow [60]	75.68	70.74	62.84	60.57	58.10	56.25
SoftFlow [25]	76.05	65.80	59.21	60.05	64.77	60.09
SetVAE [26]	76.54	67.65	55.84	60.57	59.94	59.94
DPF-Net [28]	75.18	65.55	62.00	58.53	62.35	54.48
DPM [31]	76.42	86.91	60.05	74.77	68.89	79.97
PVD [67]	73.82	64.81	56.26	53.32	**54.55**	53.83
Ours	**56.29**	**54.78**	**55.61**	**52.94**	55.75	**52.80**

Table 1. Generation results on *Airplane*, *Chair*, *Car* categories from ShapeNet [6] using 1-NN↓ as the metric.

is set as default Pair-wise Fusion Operator in experiment, where f is fully connected layers, and $\mathcal{P}$ means pair-wise alignment. We employ 3 MFM layers cascaded at the end of the denoising transformer, as filters to suppress high-frequency outliers, as validated in experiment.

4. Experiments

In this section, we quantitatively and qualitatively evaluate our proposed unified 3DQD prior on three mainstream 3D shape generation tasks. We also discuss how 3DQD is extended for different relevant applications.

4.1. 3D Shape Generation

Data and evaluation. We select three most popular categories from ShapeNet [6]: *Airplane*, *Chair*, *Car* as our main datasets for training. For input, we prepare the same T-SDF files as in DISN [58] and follow the train-test split. Following previous works [64,67], we use 1-NNA as our main metric with both Chamfer distance (CD) and earth mover distance (EMD), measuring both shape generation quality and diversity. Limitations of other metrics are also discussed in our supplementary material.

Baselines and results. We present the visual samples in Fig. 4 and quantitative results in Tab. 1. In comparison with baseline methods, we follow the same data processing and evaluation procedure as in PVD [67]. Since our data is in the format of volumetric T-SDFs, for a fair comparison, we first transform T-SDFs into meshes, and then sample 2048 points for each mesh. It is noted that 3DQD outperforms all baselines with the best shape generation quality, due to the stable iterative generation and powerful joint distribution modeling capabilities of discrete diffusion generator, while MFM layers successfully suppress high-frequency outliers and improve the smoothness of results.

4.2. Shape Completion

Data and evaluation. We evaluate our method on the ShapeNet dataset with 13 categories using the train/test splits provided by DISN [58]. As partial input is in the form of T-SDFs, we use the benchmark from AutoSDF [34], which contains two different settings of observed shapes: 1) *Bottom half* of ground truth as partial shape, and 2) *Octant*

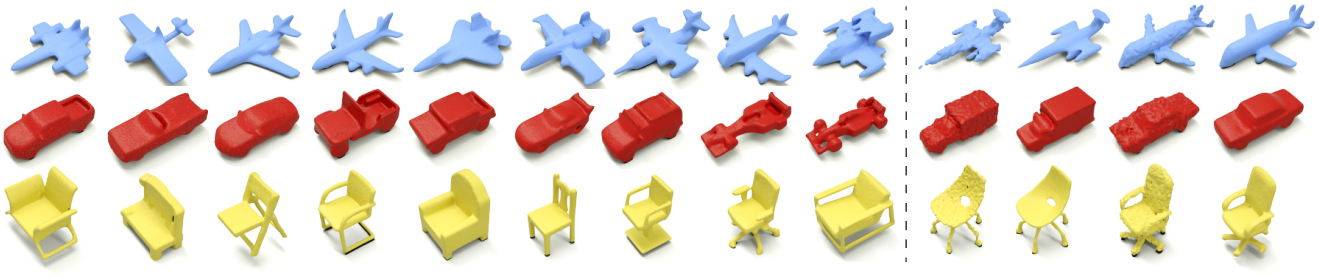

Figure 4. Our approach synthesizes high-quality and diverse shapes with smooth surfaces and complex structures on multiple tasks. *Left*: Unconditional shape generation. *Right*: Shape denoising without fine-tuning.

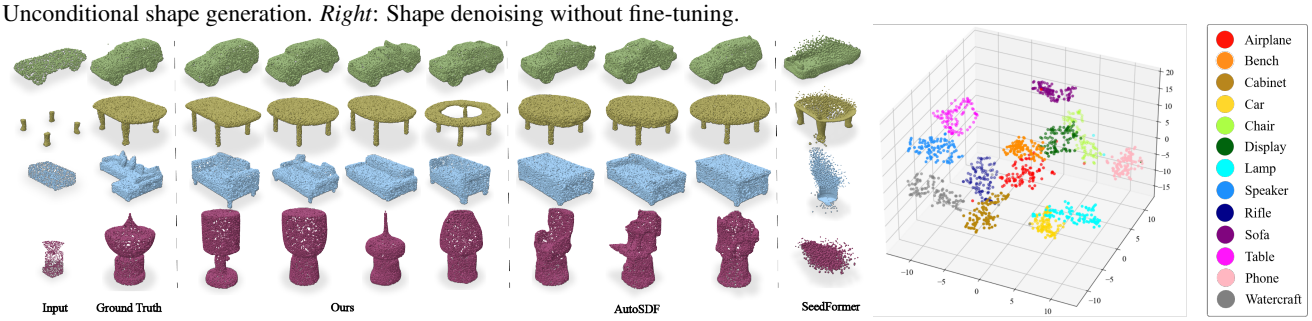

Figure 5. Shape completion comparison. Our method yields most high-fidelity and diverse generation by the powerful and iterative modeling of joint distribution from diffusion, as well as generalized prior, even though only a small part is given in the last two rows.

Figure 6. We visualize the prior with T-SNE projections. Observe intra-class coherence and inter-class divergence.

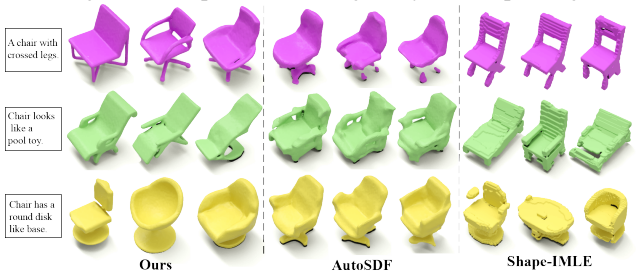

Figure 7. Comparative results for language-guided generation. Our model captures the reasonable variations from text prompts with smooth surfaces, while others produce unsatisfying results.

with front, left and bottom half of ground truth as partial shape. For evaluation, we compute *Total Mutual Difference* (TMD) of $N = 10$ generated shapes for each input. We also report *Minimum Matching Distance* (MMD) and *Average Matching Distance* (AMD) which measure the minimum and average Chamfer distance from ground truth to the N generated shapes.

Condition injection. It's vital for us to explore how to infer missing shapes in terms of the partial observations (out of distribution) with a model trained with noisy complete shape tokens. To this end, we encode the inputs into codebook tokens and diffuse them at k timestep into $\widetilde{s}_k$, to replace the full masked s_T for iterative generation. So that the reverse process starts at k timestep rather than T timestep, with partial information kept in incompletely corrupted token maps. We set $k = \frac{1}{2}T$, as smaller k damages diversity and larger k reduces the fidelity of generation shapes.

Baselines and results. We compare 3DQD with three state-of-the-art shapes completion methods, including PoinTr [62], SeedFormer [66] and AutoSDF [34]. The former two methods perform accurate supervised completion, while the latter is able to sample diverse shapes from a partial observation. A set of random held-out shapes from ShapeNet dataset with 13 categories are used to compare the performance. The number of objects in the set is $M = 500$. Quantitative results are presented in Tab. 2 and several visualized results are shown in Fig. 5. It is observed that our method surpasses all models in both diversity and fidelity of generation, thanks to the under-determined, multi-modality nature of diffusing process. It is also worth mentioning that although our model is trained only on completely noisy shape tokens, it successfully performs inference conditioned on local noisy tokens with rest regions fully masked (which it has never seen). In contrast, models which are specifically trained only for 3D completion fail this task, as illustrated in the right column of Fig. 5. It also indicates that our model has learned prior knowledge about the structural relations between local shapes, and this joint prior can be extended to more applications as discussed in Sec. 4.4.

Prior Analysis. We further visualize our learned prior distribution to provide an in-depth analysis about how the underlying shape structure distributions for various object types are well modeled by 3DQD so that they could easily facilitate various tasks. For this purpose, we sample $N = 100$ shape completion results from only an octant

Method	Bottom Half			Octant		
	MMD↓	AMD↓	TMD↑	MMD↓	AMD↓	TMD↑
PoinTr [62]	0.5316	N/A	N/A	2.1567	N/A	N/A
SeedFormer [66]	0.4972	N/A	N/A	2.3990	N/A	N/A
AutoSDF [34]	0.3510	0.8200	0.0466	0.5720	1.279	0.0826
Ours	**0.2933**	**0.6302**	**0.0478**	**0.4690**	**1.093**	**0.0960**

Table 2. Quantitative completion results on ShapeNet. MMD and AMD is multiplied by 10^2. TMD is multiplied by 10.

Method	PMMD↓	CLIP-S↑	FPD↓	TMD↑
Shape-IMLE [30]	1.681	31.42	82.34	0.0539
AutoSDF [34]	1.961	31.65	141.87	0.1302
Ours	**1.492**	**32.11**	**59.00**	**0.2795**

Table 3. Quantitative results for text-guided generation. PMMD and CLIP-S is multiplied by 10^2. TMD is multiplied by 10.

input for each object category, and then perform feature encoding for each appended completion with a pre-trained PointNet [38]. We use t-SNE [53] to project each feature representation into a point in Fig. 6. From Fig. 6, we see that 1) the visualized 3D representations completed from a same partial input are clustered, which shows that our learned prior model is able to capture common structural features for the same category; and 2) results from different inputs are obviously scattered, which demonstrates that the learned prior distribution is diverse enough (in other words, the coverage of the learned prior is sufficiently wide), so that it can differentiate different object categories. The above observations confirm that our learned prior is able to well model the underlying dependencies between local shape components, thus it serves as a general object structure probabilistic distribution, supporting a wide range of related tasks.

4.3. Language-guided Generation

Data and evaluation. We reorganize the dataset released by ShapeGlot [2] into text-shape pairs to train a text-driven conditional generative model for 3D shapes, using the train/test splits provided by AutoSDF [34]. Since most existing metrics cannot well measure the similarity between text and shape modalities, we propose new evaluation metrics for text-driven shape generation task. *CLIP-S* computes the maximum score of cosine similarity between $N = 9$ generated shapes and their text prompts by a pre-trained CLIP [39]. Since CLIP cannot handle 3D shape inputs, we render each generated shape into 20 2D images from different views to compute *CLIP-S*. In addition, We deploy *Frechet Pointcloud Distance* (FPD) and *Pairwise Minimum Matching Distance* (PMMD) to calculate the distance between ground truth and samples.

Condition injection. We first use the CLIP [39] pre-trained model (ViT-B) to encode text prompts to 77 tokens. Then parts of self-attention modules in the denoising network are replaced by cross-attention [7], for its natural compatibility with multi-modality data. As a result, the text token embeddings constantly influence the generation through cross-attention modules.

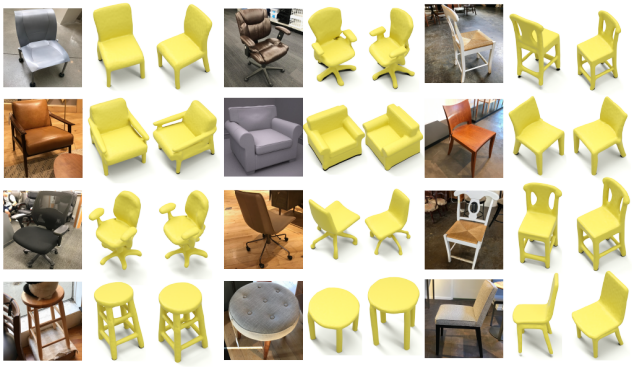

Figure 8. Single-view reconstruction results. Our model synthesizes high-quality single-view reconstruction with learned prior and cross-domain alignment.

Baselines and results. We quantitatively compare our results with two state-of-the-art methods, *i.e.*, AutoSDF [34] and Shape-IMLE [30]. We only use the generated shapes from Shape-IMLE [30] and discard their colors, to only evaluate 3D geometry quality. Quantitative results are reported in Tab. 3 and a visual comparison is shown in Fig. 7. We observe a lack of diversity from results of baselines, while our method is diversified due to randomness in diffusing process. The unified discrete local representation of codebook indexing among texts and shapes achieves good alignment between two modalities.

4.4. Extended Applications

In addition to the above major tasks, we show that 3DQD can be extended for a wide range of downstream applications, serving as a generalized prior **with little or even no tuning**. Note that in contrast, previous 3D shape priors [34, 67] DO NOT possess this generalization capability.

Denoising conditonal generation. In practice, 3D shapes captured from real scenes often have rough surfaces and noise points due to precision limitation of the capture device. It thus requires extra adaption for downstream models to accommodate noisy data, *e.g.* point clouds denoising [40]. To demonstrate our model's ability in dealing with noisy inputs, we add different levels of Gaussian and uniform noise to T-SDFs to simulate the noisy mesh surfaces in real world. Then noisy T-SDFs are encoded into shape tokens as $\widetilde{s}_k$ to replace the fully masked s_T into our pre-trained 3DQD, where $k = \frac{1}{2}T$. Then reverse Markov process starts from $\widetilde{s}_k$. Visual results are shown in Fig. 4, and quantitative evaluations with different noise levels and types are detailed in our supplementary material. From Fig. 4, it is noted that our model successfully recovers noisy inputs into clean samples **without any tuning**, demonstrating the noise immunity ability of our pre-trained 3DQD, which comes from the noise compatibility from diffusing training and quantized noise-free vectors in codebook.

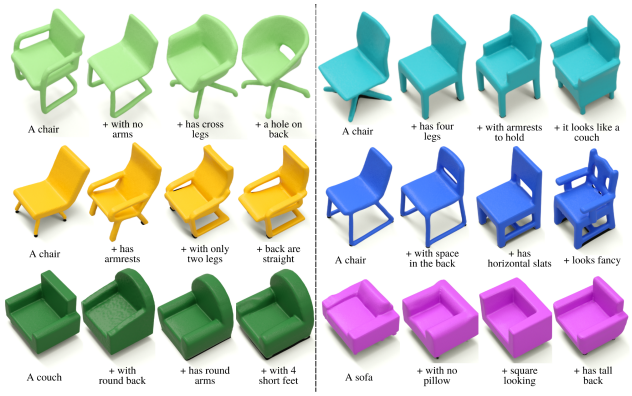

Figure 9. Text-driven shape editing results. We start from initial phrase at left and gradually edit them with new prompts. They present continuous and smooth evolution.

Method	VAE Training	Diffusion Training	Inference Time / Shape
Lion [64]	110 hours	440 hours	27.12 seconds
Ours	**9 hours**	**60 hours**	**1.61 seconds**

Table 4. Quantitative comparison of computational cost on a single V100 NVIDIA GPU. Our VAE is fine-tuned on pre-trained model released in [34] for 9 hours.

Text-driven shape editing. We also conduct an experiment on text-guided shape editing **without any tuning**. We initialize standard chair token maps with text prompts *"A chair"* with 3DQD. Then we initialize the former token maps as the start of new reverse process $\widetilde{s}_k$ for sequentially generating new shapes, with new text inputs one by one. k is set to $0.98T$ to encourage novel structure generation. The results are visualized in Fig. 9. Generated shapes are apparently high-quality and realistic, due to the structure prior memorized by diffusion generator.

Single-view reconstruction. Inferring the 3D shape from a single image is always non-trivial due to the lack of information. We show that with a well pre-trained VQ-VAE to encode and quantize images, our model achieves singe-view 3D reconstruction with a little fine-tuning. Specifically, we obtain the index coding of images with VQ-VAE released by Mittal *et al*. [34] on Pix3D dataset [49], and modify the condition embedding module of language-guided 3DQD model from discrete text tokens to discrete image tokens, with rest parts of the net remaining unchanged. We then fine-tune this conditional 3DQD model on Pix3D with masked image-shape pairs for 10 hours. For inference, we set index codes of images from VQ-VAE at the beginning of reverse process $\widetilde{s}_k$. Note that image embeddings modulate the generation process via cross-attention. Visual results are shown in Fig. 8 as accurate and realistic reconstructions with learned prior and cross-domain alignment. More results are provided in our supplementary material.

4.5. Ablation Study

4.5.1 Frequency Analysis

We experimentally analyze the frequency components of the generation to validate our design of Multi-frequency Fusion Module (MFM). We sample 1000 completion samples with three methods: our 3DQD, our 3DQD without MFM, and AutoSDF. The T-SDF results are fed to Discrete Cosine Transform to calculate its average power spectral density and obtain frequency components within each spatial volume. The power spectral density of sampled volumes is plotted in Fig. 10. It is noted that our MFM layers successfully suppress high-frequency noise and help generate shapes with smooth surfaces. More ablation studies about the architecture of MFM is in our supplementary material.

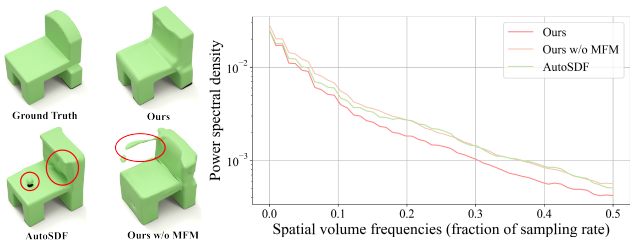

Figure 10. Power spectral density of different frequencies on 1000 generated shapes along with visual comparison. Our MFM layers successfully suppress high-frequency components.

4.5.2 Efficiency Analysis

We compare our model with the latest 3D diffusion work Lion [64] (with no open source codes) about training and inference complexity, evaluated on single-class shape generation task as shown in Tab. 4. It is observed that our model only requires 69 (9+60) hours for training totally while Lion [64] spends as much as 550 (110+440) hours. Also note that our reported results are all generated with 100-step DDPM-based sampling, much more efficient than that of Lion (*i.e*. 1000 steps), with visually competitive performance compared to Lion [64].

5. Conclusion

We have introduced 3DQD, a unified and efficient shape generation prior model for multiple 3D tasks. Our model first learns a compact representation with P-VQ-VAE for its advantages in computational saving and consistency among different tasks. Then a novel discrete diffusion generator is trained with accurate, expressive and diversified object structural modeling. Multi-frequency fusion modules are developed to suppress high-frequency outliers. Solid experiments and rich analysis have demonstrated our approach possesses superior generative power and impressive shape generation quality on various 3D shape generation tasks. Furthermore, our prior model can serve as a generalized backbone for multiple downstream tasks with no or little tuning, while no architectural change is needed.

6. Acknowledge

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 5

[2] Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. Shapeglot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8938–8947, 2019. 7

[3] Oron Ashual, Shelly Sheynin, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 2

[4] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021. 2, 5

[5] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *European Conference on Computer Vision*, pages 364–381. Springer, 2020. 2

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[7] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 5, 7

[8] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, pages 100–116. Springer, 2018. 2

[9] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 2

[10] Xuelin Chen, Baoquan Chen, and Niloy J Mitra. Unpaired point cloud completion on real scans using adversarial training. *arXiv preprint arXiv:1904.00069*, 2019. 2

[11] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3, 5

[14] Rao Fu, Xiao Zhan, Yiwen Chen, Daniel Ritchie, and Srinath Sridhar. Shapecrafter: A recursive text-conditioned 3d shape generation model. *arXiv preprint arXiv:2207.09446*, 2022. 1, 2

[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2, 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 4, 5

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[21] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021. 2, 4

[22] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. *arXiv preprint arXiv:2209.08725*, 2022. 1

[23] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1251–1261, 2020. 3

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[25] Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. Softflow: Probabilistic framework for normalizing flow on manifolds. *Advances in Neural Information Processing Systems*, 33:16388–16397, 2020. 5

[26] Jinwoo Kim, Jaehoon Yoo, Juho Lee, and Seunghoon Hong. Setvae: Learning hierarchical composition for generative modeling of set-structured data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15059–15068, 2021. 5

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[28] Roman Klokov, Edmond Boyer, and Jakob Verbeek. Discrete point flow networks for efficient point cloud generation. In

European Conference on Computer Vision, pages 694–710. Springer, 2020. 5

[29] Ruihui Li, Xianzhi Li, Ka-Hei Hui, and Chi-Wing Fu. Sp-gan: Sphere-guided 3d shape generation and manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 1, 2

[30] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2022. 1, 2, 7

[31] 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. 1, 2, 3, 5

[32] Priyanka Mandikal, KL Navaneet, Mayank Agarwal, and R Venkatesh Babu. 3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. *arXiv preprint arXiv:1807.07796*, 2018. 2

[33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[34] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 1, 2, 3, 5, 6, 7, 8

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 5

[36] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[37] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 7

[39] 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. 7

[40] Marie-Julie Rakotosaona, Vittorio La Barbera, Paul Guerrero, Niloy J Mitra, and Maks Ovsjanikov. Pointcleannet: Learning to denoise and remove outliers from dense point clouds. In *Computer Graphics Forum*, 2020. 7

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

[42] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2

[44] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[45] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[46] Marcelo dos Santos, Rayson Laroca, Rafael O Ribeiro, João Neves, Hugo Proença, and David Menotti. Face super-resolution using stochastic differential equations. *arXiv preprint arXiv:2209.12064*, 2022. 2

[47] Yue Shi, Bingbing Ni, Jinxian Liu, Dingyi Rong, Ye Qian, and Wenjun Zhang. Geometric granularity aware pixel-to-mesh. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13097–13106, 2021. 2

[48] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3859–3868, 2019. 1

[49] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 8

[50] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 2, 4

[51] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 383–392, 2019. 2

[52] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3, 5

[53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[54] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the Euro-*

pean conference on computer vision (ECCV), pages 52–67, 2018. 1, 2

[55] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 3dn: 3d deformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1038–1046, 2019. 2

[56] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018. 2

[57] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 829–838, 2020. 2

[58] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 5

[59] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022. 3

[60] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019. 1, 2, 5

[61] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 2

[62] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12498–12507, 2021. 1, 2, 6, 7

[63] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018. 2

[64] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 2, 3, 4, 5, 8

[65] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1768–1777, 2021. 2

[66] Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang. Seedformer: Patch seeds based point cloud completion with upsample transformer. *arXiv preprint arXiv:2207.10315*, 2022. 1, 2, 6, 7

[67] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 1, 2, 3, 4, 5, 7