Dora: Sampling and Benchmarking for 3D Shape Variational Auto-Encoders

Rui Chen[1,2] Jianfeng Zhang[2†] Yixun Liang[1,3] Guan Luo[2,4] Weiyu Li[1,3]
Jiarui Liu[1,3] Xiu Li[2] Xiaoxiao Long[1,3] Jiashi Feng[2] Ping Tan[1,3†]
[†]Corresponding authors
[1]The Hong Kong University of Science and Technology [2]ByteDance Seed
[3]LightIllusions [4]Tsinghua University
https://aruichen.github.io/Dora

Abstract

Recent 3D content generation pipelines commonly employ Variational Autoencoders (VAEs) to encode shapes into compact latent representations for diffusion-based generation. However, the widely adopted uniform point sampling strategy in Shape VAE training often leads to a significant loss of geometric details, limiting the quality of shape reconstruction and downstream generation tasks. We present Dora-VAE, a novel approach that enhances VAE reconstruction through our proposed sharp edge sampling strategy and a dual cross-attention mechanism. By identifying and prioritizing regions with high geometric complexity during training, our method significantly improves the preservation of fine-grained shape features. Such sampling strategy and the dual attention mechanism enable the VAE to focus on crucial geometric details that are typically missed by uniform sampling approaches. To systematically evaluate VAE reconstruction quality, we additionally propose Dora-bench, a benchmark that quantifies shape complexity through the density of sharp edges, introducing a new metric focused on reconstruction accuracy at these salient geometric features. Extensive experiments on the Dora-bench demonstrate that Dora-VAE achieves comparable reconstruction quality to the state-of-the-art dense XCube-VAE while requiring a latent space at least 8× smaller (1,280 vs. > 10,000 codes). Project page: https://aruichen.github.io/Dora.

1. Introduction

3D content creation is vital to delivering realistic and immersive experiences in various industries, including games, movies, and AR/VR. However, traditional 3D modeling typically demands significant expertise and manual effort, making it time-consuming and challenging, especially for non-expert users. Recent advances in AI-powered 3D content generation methods [2, 3, 7, 9, 17, 21, 25, 28, 30, 42, 44, 57] have transformed the field, making it more accessible to many more users.

Following the success of text-to-image generation models [4, 5, 15, 37, 55], recent 3D content creation approaches [36, 49, 56] adopt a two-stage pipeline: encoding 3D shapes into a latent space using variational autoencoders (VAEs), followed by training a latent diffusion model. The performance of such a generative pipeline heavily relies on the VAE's capability to faithfully encode and reconstruct 3D shapes.

Existing 3D VAEs operate by sampling points on mesh surfaces for shape encoding and then reconstructing the original 3D meshes by its decoder. This process faces unique challenges compared to 2D image VAEs where the input image is fully observable. In comparison, the sampled point cloud often cannot capture all the necessary shape information, which could harm the performance of 3D VAEs.

Volume-based method [36] leverages sparse convolution [47] to process millions of voxelized points for high-fidelity reconstruction. Its dense sampling captures precise shape information. However, this method produces large latent codes (commonly $>$ 10,000 tokens), which significantly complicate the training of diffusion models. On the other hand, vector-set (Vecset) methods [24, 54, 56, 58] use transformers to achieve compact latent representations (hundreds to thousands of tokens), enabling efficient diffusion [24, 56]. However, due to the quadratic complexity of transformer networks, it often only samples a few thousand points to represent a 3D shape, which leads to information loss and performance degradation. Therefore, we seek to improve the reconstruction quality of Vecset-based VAEs and maintain their compact representation.

We begin by analyzing the shape reconstruction capability of Vecset-based VAEs. Through careful analysis, we find these methods have limited reconstruction performance, which stems from their commonly used uniform sampling. When computational constraints limit the total

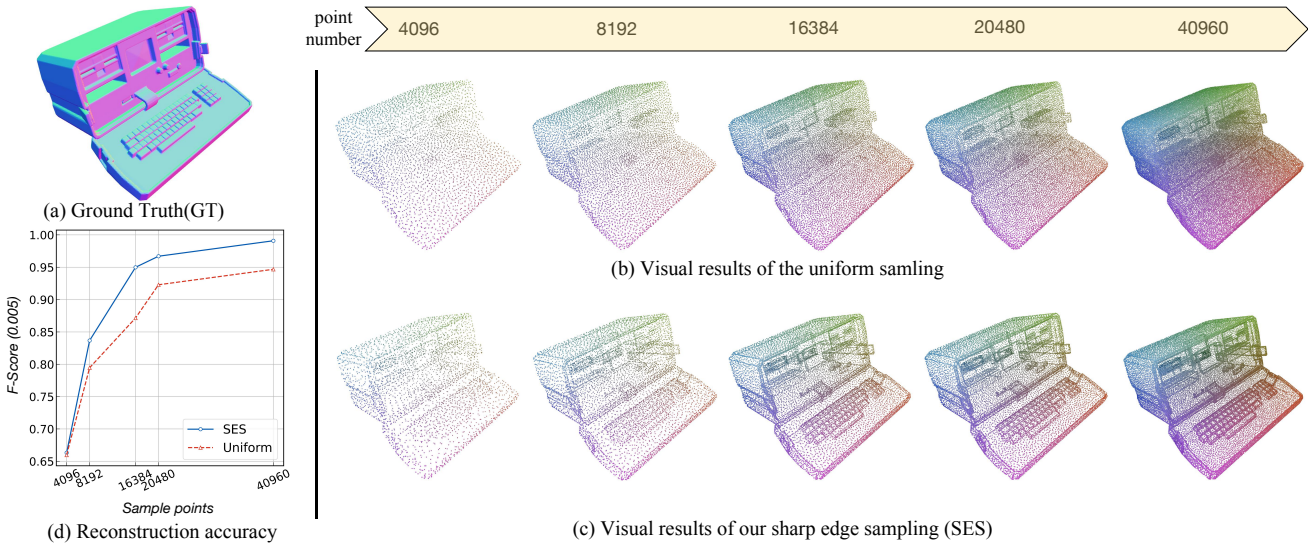


Figure 1. Sampling strategy comparison. Given the ground truth mesh shown in (a), we visualize point clouds produced by uniform sampling in (b) and those generated by our proposed Sharp Edge Sampling (SES) in (c), at various sampling rates. In (d), we compare the reconstruction accuracy trained with our SES and the uniform sampling using the F-score metric. The comparison demonstrates SES consistently outperforms uniform sampling under varying sampling rates, as the point clouds generated by SES are more effective in capturing the salient features of the object.

number of sampled points, uniform sampling fails to prioritize geometrically salient regions, leading to the loss of fine details. To validate this observation, we experiment on a 3D mesh with intricate geometric details (e.g., keyboard buttons) shown in Figure 1 (a). We visualize the point cloud with different sampling strategies at various sampling densities in (b) and (c). As demonstrated in Figure 1 (b), even with increasing sampling rates, uniform sampling fails to preserve sharp features like keyboard buttons. This simple experiment confirms that uniform sampling fundamentally limits the capturing of geometric details, which in turn affects the VAE's reconstruction capability and the result quality of the learned diffusion models.

Inspired by the success of importance sampling in various geometric processing tasks [14, 48], we introduce a similar strategy for 3D VAE training. While existing importance sampling methods focus on down-sampling point clouds, our task requires sampling points directly from mesh surfaces for shape VAE training [49, 56]. This fundamental difference necessitates a new sampling approach specifically designed for preserving geometric features from mesh representations.

To address this challenge, we propose a **Sharp Edge Sampling (SES)** algorithm that adaptively samples points based on geometric saliency. Specifically, SES first identifies edges with significant dihedral angles on the mesh, indicating regions of high geometric complexity. It then samples points along these salient regions while maintaining a balance with uniformly sampled points to capture the overall structure. This approach ensures comprehensive coverage of both fine details and global geometry.

Building upon SES, we present Dora-VAE, a novel method achieving high-fidelity reconstruction while maintaining compact latent representations. To fully leverage these detail-rich point clouds sampled by our SES, we design a dual cross-attention architecture that effectively processes both salient and uniform regions during encoding. As shown in Figure 1, our method significantly outperforms uniform sampling in preserving shape details on the keyboard (c), with consistent improvements in F-Score across different sampling rates (d).

The common evaluation protocol for 3D VAEs is also biased. It typically uses a set of randomly selected 3D shapes, and employs general metrics (e.g., F-score, Chamfer distance) to measure the shape reconstruction quality. However, we argue it is necessary to divide the test shapes into sub-classes of different shape complexity to better evaluate these 3D VAEs. To facilitate the evaluation of 3D VAEs, we further introduce Dora-Bench, a new benchmark based on existing datasets with our novel Sharp Normal Error (SNE) metric. Dora-Bench categorizes test shapes based on their geometric complexity and SNE focuses on measuring the reconstruction quality of salient geometric features. This combination enables a more rigorous assessment of 3D VAEs.

Extensive experiments demonstrate that our Dora-VAE achieves superior results. When integrated into downstream 3D diffusion models, it significantly enhances the qual-

ity of generated 3D shapes, which validates that our novel sampling strategy and dual cross-attention architecture effectively preserve geometric details with compact latent spaces. To summarize, our main contributions include: 1) We propose Dora-VAE, a novel 3D VAE model for high-quality reconstruction with compact latent representations, accompanied by Dora-Bench, a comprehensive benchmark for evaluating 3D VAEs. 2) We introduce, for the first time, importance sampling to the task of 3D VAE learning and propose a Sharp Edge Sampling (SES) algorithm to prioritize geometrically salient regions. Building on SES, we design a dual cross-attention architecture to effectively encode these detail-rich point clouds. 3) We develop a systematic evaluation benchmark based on existing datasets with our novel Sharp Normal Error (SNE) metric that specifically assesses reconstruction accuracy of fine geometric details, enabling more rigorous evaluation than conventional random sampling approaches.

2. Related work

Importance Sampling in Point Clouds. Importance sampling techniques have been widely used in point cloud processing tasks [14, 48]. For instance, APES [48] proposes attention-based sampling for point cloud classification and segmentation. However, these methods operate directly on point clouds rather than meshes, making them less suitable for VAE-based shape representation where preserving complete geometric information is crucial.

3D Shape VAEs. Recent 3D shape VAEs follow two main approaches: volume-based and vector set-based. Volume-based methods [36, 50] like XCube [36] use sparse convolution to encode voxelized surfaces, achieving high reconstruction quality but requiring large latent codes ($>$ 10,000 tokens). While these methods excel at preserving geometric details, their large latent spaces pose significant challenges for downstream diffusion model training. In contrast, vector set-based approaches [24, 49, 54, 56, 58] encode uniformly sampled surface points using transformers, producing highly compact latent spaces that are particularly suitable for diffusion models. However, these methods often struggle with geometric detail preservation, especially in regions with complex surface features. Our analysis reveals that this limitation primarily stems from their uniform sampling strategy: when computational constraints restrict the total number of processable points, uniform sampling fails to prioritize geometrically significant regions, leading to insufficient capture of fine details. This information loss at the sampling stage fundamentally limits these methods' ability to learn and preserve intricate geometric features.

3D Content Creation. Current 3D generation methods can be categorized into three groups. Optimization-based methods [6, 23, 26, 27, 30, 34, 35, 39, 41, 45], pioneered by DreamFusion [34] utilize score distillation sampling (SDS) to optimize 3D representations [19, 31, 38] using 2D diffusion model priors. While these methods can achieve photorealistic results, they suffer from slow generation speed, training instability, and often struggle to maintain geometric consistency. Large reconstruction models, like LRM [16] and follow-up works [22, 29, 40, 43, 46, 51] employ large-scale sparse-view reconstruction for efficient 3D generation. However, their lack of explicit geometric priors often leads to compromised geometric fidelity and inconsistent surface details. 3D native generative models [24, 49, 54, 56, 58], represented by 3DShape2VecSet [54], adopt a two-stage approach: first training a 3D VAE to encode shapes into latent space, then training a conditional latent diffusion model for generation. This approach ensures better geometric consistency through the VAE's built-in geometric constraints. However, the quality of generated shapes is fundamentally limited by the VAE's reconstruction capability. Recent works [24, 53] have shown that improving VAE reconstruction directly enhances downstream generation quality, which motivates our focus on advancing VAE design.

3. Method

In this section, we present **Dora-VAE** for high-quality 3D reconstruction, and **Dora-Bench** for 3D VAE evaluation. We first briefly review 3DShape2VecSet [54] in Section 3.1, the foundation of our method and then detail our key innovations in Section 3.2 and Section 3.3.

3.1. Preliminary: 3DShape2VecSet

3DShape2VecSet [54] introduces a transformer-based 3D VAE that encodes uniformly sampled surface points into compact latent codes. Given a 3D surface S, their pipeline consists of three key steps:

- Surface Sampling: Uniformly sample N_d points on the surface S using Poisson disk sampling [52] to obtain a dense point cloud P_d, then downsample it to N_s points via Farthest Point Sampling (FPS) [32] to get a sparse point cloud P_s:

$$P_d = \{p_d^i \in S \mid i = 1, ..., N_d\}, P_s = \text{FPS}(P_d, N_s). \quad (1)$$

- Feature Encoding: Compute the point cloud feature C via the cross-attention between P_s and P_d, followed by some self-attention layers to generate the latent code z:

$$C = \text{CrossAttn}(P_s, P_d, P_d), z = \text{SelfAttn}(C). \quad (2)$$

- Geometry Decoding: Further decode z through self-attention layers and predict occupancy values using randomly sampled spatial query points $Q_{space} \in \mathbb{R}^3$:

$$\hat{O} = \text{CrossAttn}(Q_{space}, \text{SelfAttn}(z)). \quad (3)$$

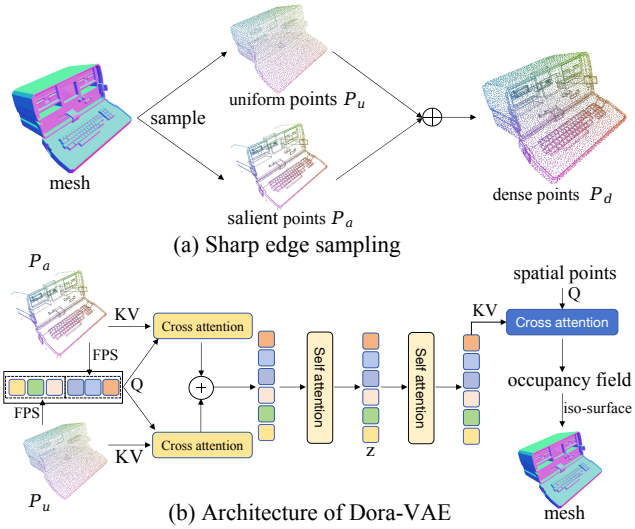


Figure 2. **Overview of Dora-VAE.** (a) We utilize the proposed sharp edge sampling technique to extract both salient and uniform points from the input mesh. These points are then combined with dense points, effectively capturing both salient regions and smooth areas. (b) To enhance the encoding of point clouds sampled through sharp edge sampling, we design a dual cross-attention architecture.

While this method generates compact latent codes for diffusion, the uniform sampling limits its ability to capture fine geometric details. Our work addresses this limitation by carefully designed sampling and encoding strategies.

3.2. Dora-VAE

Figure 2 gives an overview of our pipeline. For each input mesh, we augment the uniformly sampled point cloud P_u with more important points P_a sampled by our proposed sharp edge sampling strategy, which forms the dense point cloud P_d. During the encoding process, we compute the attention for P_u and P_a separately via a simple-yet-effective dual cross attention mechanism and sum the results for self-attention to compute the latent code z. Our VAE training largely follows that of 3DShape2VecSet [54], which is supervised by a loss evaluated on the occupancy field.

3.2.1. Sharp Edge Sampling (SES)

We propose SES to effectively sample points from geometrically salient regions. To ensure surface coverage, we also sample points uniformly. Our final sampled point cloud P_d combines uniformly sampled points P_u with points specifically sampled from salient regions P_a as $P_d = P_u \cup P_a$. Our method computes salient points P_a through two steps: detecting salient edges and sampling points from these regions.

Salient Edges Detection. Given a triangular mesh, we identify a set of salient edges Γ by analyzing dihedral angles between adjacent faces, which calculates the angle between the normal vectors of adjacent faces, providing a direct measure of surface curvature at mesh edges. For each edge e shared by adjacent faces f_1 and f_2, we compute the dihedral angle θ_e as:

$$\theta_e = \arccos\left(\frac{\mathbf{n}_{f_1} \cdot \mathbf{n}_{f_2}}{\|\mathbf{n}_{f_1}\|\|\mathbf{n}_{f_2}\|}\right), \tag{4}$$

where $\mathbf{n}_{f_1}$ and $\mathbf{n}_{f_2}$ are the normals of f_1 and f_2. The salient edge set Γ contains all edges with a dihedral angle exceeding a predefined threshold τ:

$$\Gamma = \{e \mid \theta_e > \tau\} \tag{5}$$

Let $N_\Gamma = |\Gamma|$ represent the number of the salient edges.

Salient Points Sampling. For each salient edge $e \in \Gamma$, we collect its two vertices $v_{e,1}$ and $v_{e,2}$ into a salient vertex set P_Γ:

$$P_\Gamma = \{v_{e,1}, v_{e,2} \mid e \in \Gamma\}, \tag{6}$$

where duplicate vertices from connecting edges are included only once. Let $N_V = |P_\Gamma|$ denote the number of unique vertices in P_Γ.

Given a target number of salient points N_{desired}, we generate the salient point set P_a based on the available salient vertices:

$$P_a = \begin{cases} \text{FPS}(P_\Gamma, N_{\text{desired}}), & \text{if } N_{\text{desired}} <= N_V, \\ P_\Gamma \cup P_{\text{interpolated}}, & \text{if } 0 < N_V < N_{\text{desired}}, \\ \emptyset, & \text{if } N_V = 0. \end{cases} \tag{7}$$

When we have excess salient vertices ($N_{\text{desired}} \leq N_V$), we use FPS to downsample P_Γ to obtain P_a. For cases with insufficient salient vertices ($N_V < N_{\text{desired}}$), we include all vertices from P_Γ and supplement with additional points $P_{\text{interpolated}}$. These additional points are generated by uniformly sampling $(N_{\text{desired}} - N_V)/N_\Gamma$ points along each salient edge in Γ, ensuring comprehensive coverage of salient features. When no salient edges are detected ($N_V = 0$), P_a remains empty.

3.2.2. Dual Cross Attention

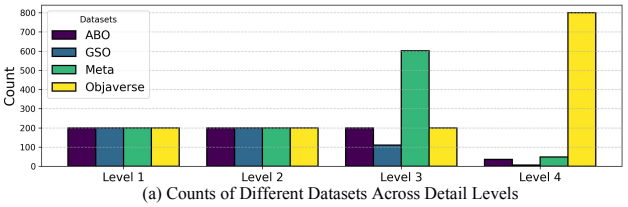
Given the point clouds P_d produced by our SES strategy, we design a dual cross-attention architecture to effectively encode both uniform and salient regions. Following 3DShape2VecSet [54], we first downsample P_u and P_a seperately using FPS:

$$P_s = \text{FPS}(P_u, N_{s,1}) \cup \text{FPS}(P_a, N_{s,2}), \tag{8}$$

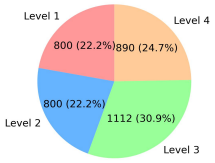
where $N_{s,1}$ and $N_{s,2}$ is the number of downsampled point clouds from P_u and P_a, respectively. We then compute cross-attention features separately for uniform and salient points as follows:

$$C_u = \text{CrossAttn}(P_s, P_u, P_u) \tag{9}$$

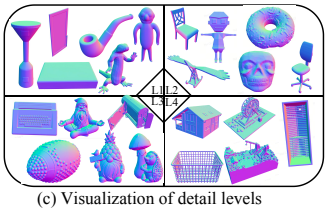
$$C_a = \text{CrossAttn}(P_s, P_a, P_a) \tag{10}$$



(a) Counts of Different Datasets Across Detail Levels



(b) Total Counts by Detail Levels



(c) Visualization of detail levels

Figure 3. Our proposed benchmark include 3D shapes from the ABO [10], GSO [13], Meta [1], and Objaverse [12] datasets. (a) The histogram of different datasets across different shape complexities. (b) The pie chart of the total counts by shape complexities. (c) Sample shapes of different shape complexities.

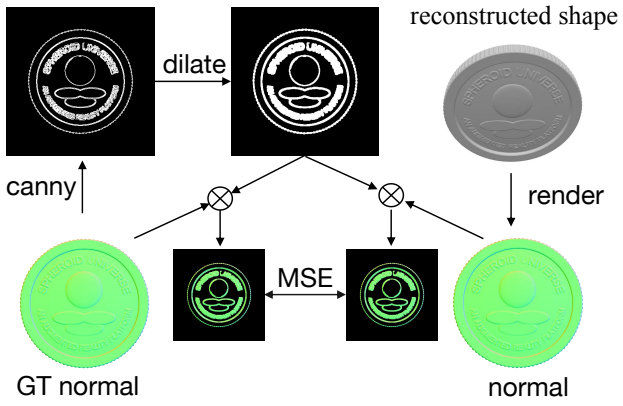


Figure 4. The process of computing sharp normal errors (SNE). We compute MSE loss in the sharp regions of the normal.

The final point cloud feature C combines both attention results:

$$C = C_u + C_a. \tag{11}$$

This dual attention design enables separate focus on uniform and salient regions during feature extraction. Following 3DShape2VecSet [54], we use C to predict the occupancy field $\hat{O}$ through self-attention blocks. The whole model with parameters ψ are optimized using MSE loss:

$$\nabla_\psi \mathcal{L}_{\text{MSE}}(\hat{O}, O) = \mathbb{E}\left[2(\hat{O} - O)\frac{\partial \hat{O}}{\partial \psi}\right]. \tag{12}$$

3.3. Dora-Bench

3.3.1. Geometric Complexity-based Evaluation

To enable a more rigorous evaluation of VAE performance, we propose Dora-bench, a benchmark that systematically categorizes test shapes based on their geometric complexity. Unlike previous methods that use randomly selected test sets, we measure shape complexity using the number of salient edges N_Γ (Section 3.2.1) and classify shapes into four levels:

- **Level 1 (Less Detail)**: $0 < N_\Gamma \leq 5000$;
- **Level 2 (Moderate Detail)**: $5000 < N_\Gamma \leq 10000$;
- **Level 3 (Rich Detail)**: $10000 < N_\Gamma \leq 50000$;
- **Level 4 (Very Rich Detail)**: $N_\Gamma > 50000$.

We curate test shapes from multiple public datasets including GSO [13], ABO [10], Meta [1], and Objaverse [12] to ensure diverse geometric complexities. Figure 3 shows the distribution of shapes across complexity levels (a,b) and example meshes from each level (c). Please refer to our supplementary materials for more examples.

3.3.2. Sharp Normal Error (SNE)

Building on our Dora-bench, we further introduce Sharp Normal Error (SNE) to evaluate reconstruction quality in salient regions. While existing metrics like Chamfer Distance and F-Score capture overall shape similarity, they fail to specifically assess the preservation of fine geometric details. SNE addresses this limitation by measuring normal map differences between reconstructed and ground truth shapes in geometrically significant areas. As illustrated in Figure 4, we render normal maps of the ground truth shape from multiple viewpoints and identify salient regions using Canny edge detection. These regions are dilated to create evaluation masks. The final SNE metric is computed as the Mean Squared Error between ground truth and reconstructed normal maps within the masked areas. This process enables focused evaluation of how well VAEs preserve sharp geometric features during reconstruction.

4. Experiments

We conducted intensive experiments to validate the effectiveness of our proposed 3D VAE and compare it with other state-of-the-art methods.

4.1. Implementation Details

We follow CLAY [56] for mesh preprocessing to ensure watertight 3D models. Our VAE is trained on a subset filtered from Objaverse [12], containing approximately 400,000 3D meshes. We filter out low-quality meshes with missing faces or severe self-intersections to ensure training stability. Our training is conducted on 32 A100 GPUs for two days using a batch size of 2048 and a learning rate of 5e-5. We employ Flash-Attention-v2 [11], mixed-precision training with FP16 and gradient checkpointing [8] to optimize

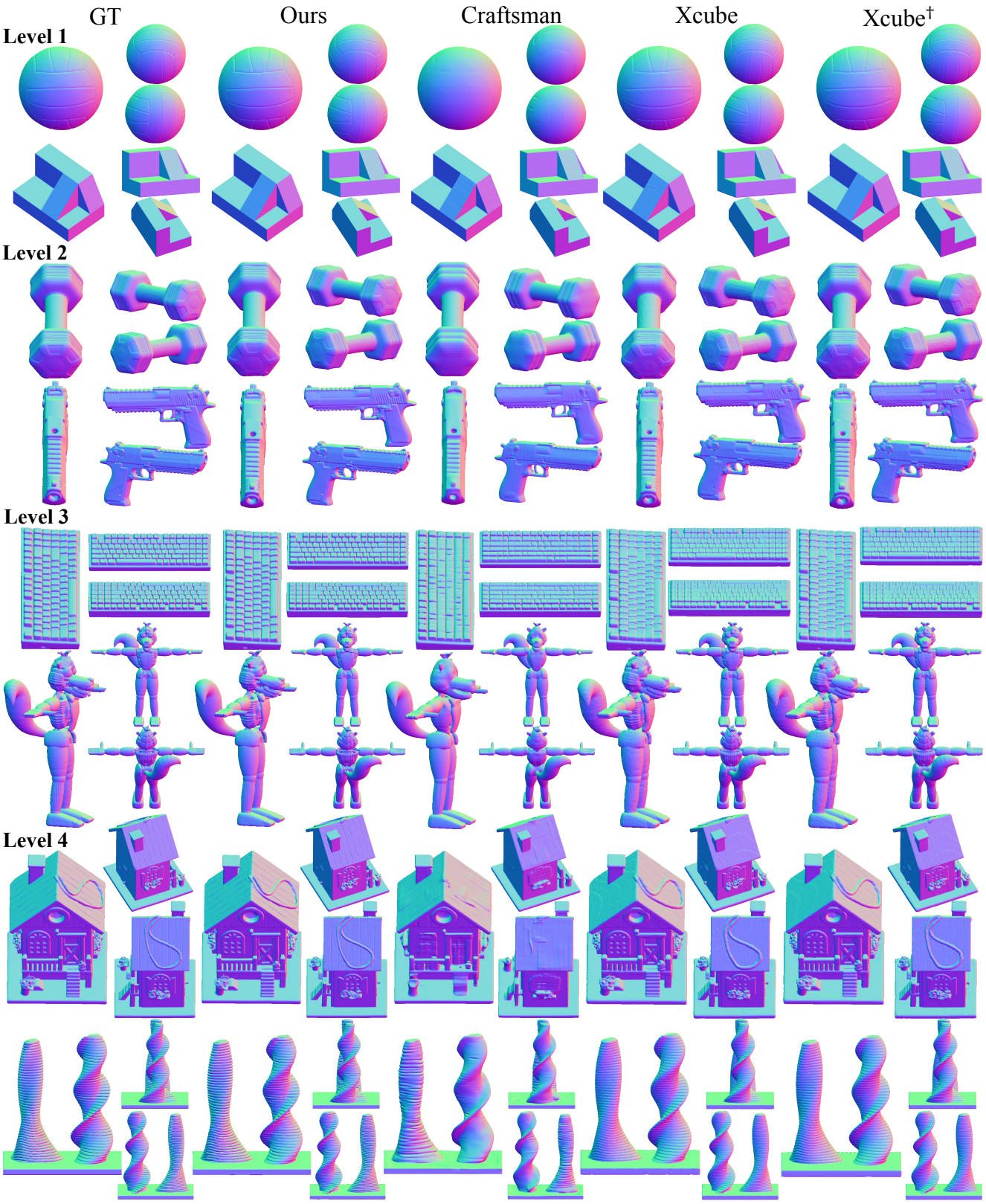


Figure 5. Qualitative comparison of the VAE reconstruction results. † indicates the fine-tuning model that uses the same training data as ours.

Methods	LCL	↑ F-score(0.01) × 100				↑ F-score(0.005) × 100				↓ CD × 10000				↓ SNE × 100			
		L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4
Xcube [36]	>10000	98.968	98.799	98.615	98.226	95.525	93.872	92.322	85.365	6.315	6.288	7.935	9.926	1.579	1.432	1.430	1.679
Xcube† [36]	>10000	99.393	99.794	99.824	99.079	96.753	95.535	93.422	87.365	4.015	4.142	5.740	7.627	1.543	1.408	1.259	1.639
Craftsman [24]	256	98.016	95.874	91.756	81.739	87.994	82.549	73.000	57.379	4.389	9.129	14.530	33.441	1.906	1.873	2.191	3.933
Ours w/o DCA	1280	99.964	99.925	99.678	97.890	96.561	95.975	91.618	83.124	2.236	2.506	4.444	6.432	1.448	1.215	1.205	1.828
Ours w/o SES,DCA	1280	99.944	99.814	97.294	96.779	95.977	94.623	88.406	79.240	2.422	2.983	3.980	6.196	1.496	1.313	1.352	2.207
Ours full	256	99.507	98.986	96.669	89.577	93.272	90.466	82.386	68.669	3.356	5.202	10.276	24.527	1.555	1.410	1.618	3.035
	1280	**99.988**	**99.955**	**99.880**	**99.170**	**97.038**	**96.831**	**93.458**	**87.473**	**2.097**	**2.500**	**3.945**	**5.265**	**1.433**	**1.186**	**1.137**	**1.579**

Table 1. Quantitative comparison in Dora-bench. † indicates the fine-tuning model that uses the same training data as ours.

memory usage and training efficiency. For the parameters in sharp edge sampling, we set $N_{\text{desired}} = 16384$ and $\tau = 30$. We set the low threshold to 20 and the high threshold to 200 for the canny edge detection.

4.2. Evaluation Setting

Metrics. We evaluate reconstruction quality by comparing input meshes with their decoded counterparts from different 3D VAEs using 1M sampled points under three metrics: 1) F-score (r) [20], which reconstruction accuracy by computing precision and recall of point correspondences within distance threshold r. Specifically, we report F-score (0.01) and F-score (0.005) with shapes normalized $[-1, 1]$. 2) Chamfer Distance (CD), which computes the average distance between each reconstructed point and its nearest ground truth point. 3) Sharp Normal Errors (SNE) as proposed in Section 3.3.2, which evaluate normal map differences in salient regions. For fair comparison, we also report Latent Code Length (LCL) as longer codes typically enable better reconstruction.

Baselines. We compare Dora-VAE with state-of-the-art approaches, including: 1) XCube-VAE [36], a volumetric method with larger latent codes; 2) XCube-VAE† [36], our fine-tuned version of the original XCube-VAE on the same dataset; 3) Craftsman-VAE[24], which fine-tune the 3DShape2VecSet [54] with shorter latent codes on Objaverse. We exclude VAE models from Direct3D [49] and CLAY[56] as as their implementations were not publicly available at submission time.

4.3. Qualitative Comparison

Figure 5 shows visual comparisons of different methods across different complexity levels from our Dora-bench dataset. We visualize both ground truth and reconstructed meshes using surface normal coloring to highlight geometric details. For shapes with lower complexity (L1 and L2), all methods achieve comparable reconstruction quality. However, when dealing with shapes of higher complexity (L3 and L4), the advantages of our method become obvious. While XCube-VAE achieves similar visual quality to ours, it requires a significantly larger latent space - more than 8× the size of ours (> 10,000 dimensions vs. 1,280). This substantial reduction in latent code length, while maintaining high reconstruction fidelity, makes our method particularly suitable for training 3D diffusion models. In contrast, Craftsman-VAE shows a noticeable degradation in reconstruction quality for complex shapes, failing to capture fine geometric details. Additional visual comparisons are provided in the appendix.

4.4. Quantitative Comparison

Table 1 presents quantitative results of different methods across different complexity levels of Dora-bench. Our method consistently outperforms baselines across all levels, with larger margins on more complex shapes (L3 and L4). The advantage is particularly evident in CD metrics, where our method with only 256 latent codes surpasses even our fine-tuned version of XCube-VAE (3.356 vs. 4.015). When using 1280 latent codes, our method further decreases CD to 2.097, achieving a 47.77% improvement over XCube-VAE†. We attribute XCube-VAE's lower performance to its use of NKSR [18] for mesh extraction, which introduces additional quantization errors.

Notably, our method demonstrates superior performance in preserving geometric details, as reflected by the SNE metric. For example, in L4 shapes where geometric complexity is highest, our method achieves an SNE of 1.579 compared to 1.639 from XCube-VAE†, representing a 3.7% improvement. This significant gain in SNE aligns with our qualitative observations in Figure 5, where our method better preserves fine details such as sharp edges and complex surface variations, demonstrating the effectiveness of our sharp edge sampling strategy.

4.5. Ablation Studies

To evaluate the contribution of each component, we compare our full model with two variants under the same training conditions:

- Ours w/o SES, DCA. This variant removes both sharp edge sampling (SES) and dual cross attention (DCA), i.e.

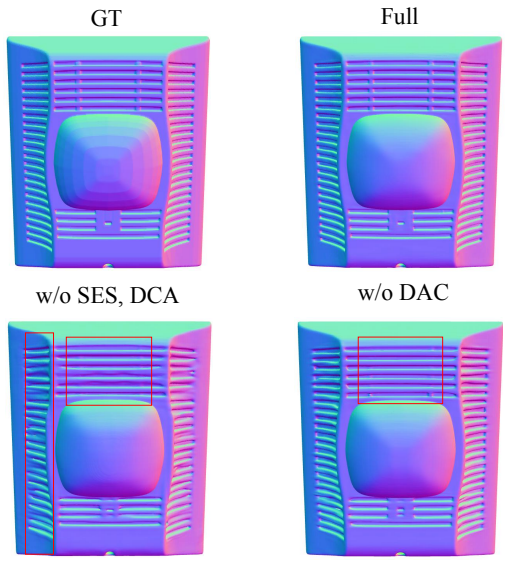


Figure 6. Ablation studies of our method. Given the ground truth of mesh, we employ both our full model and its variations to reconstruct the ground truth mesh, highlighting significant reconstruction discrepancies with red boxes.

using only uniformly sampled point clouds with Poisson disk sampling [52], while maintaining an equal N_d.

- Ours w/o DCA. This variant retains SES but removes DCA, i.e. using a single cross attention adopted by [54].

As shown in Figure 6 and Table 1, our full model consistently outperforms these variants, validating the effectiveness of both components.

5. Application: Single Image to 3D

We demonstrate the effectiveness of our VAE by applying it to single-image 3D generation through diffusion models. Following CLAY [56], we implement a latent diffusion model based on the DiT [33] architecture. For a fair comparison, we fine-tune Craftsman-VAE on our dataset (denoted as Craftsman-VAE†) since both Craftsman-VAE and 3DShape2VecSet were originally trained on smaller datasets. Note that XCube-VAE is excluded from comparison due to its 10,000-dimensional latent codes being impractical for diffusion model training. Figure 7 shows some generation results from diffusion models trained with our Dora-VAE and Craftsman-VAE†. Both models share identical architecture (0.39B parameters) and training conditions (same dataset, 32 A100 GPUs, 3 days). Our Dora-VAE demonstrates significantly better preservation of geometric details in the generated shapes, validating its effectiveness as a foundation for 3D generation tasks. While further improvements could be achieved with more extensive training data and computational resources, we focus on validating the VAE's capabilities in this work and leave such extensions for future work.

Figure 7. The diffusion results of the single image to 3D generation trained on our Dora-VAE and Craftsman†. The 3D geometry generated by the diffusion model trained on our proposed Dora-VAE has more details under the same experimental environment.

6. Conclusion

In this work, we introduce Dora-VAE, a novel VAE designed for high-quality 3D shape compression and reconstruction. At its core, Dora-VAE introduces sharp edge sampling to effectively capture salient geometric features, complemented by a dual cross-attention architecture that enhances the encoding of these detail-rich point clouds. To enable more rigorous evaluation of VAE performance, we develop Dora-bench, which systematically categorizes shapes based on geometric complexity and introduces the Sharp Normal Error (SNE) metric for specifically assessing the preservation of fine geometric details. Our comprehensive experiments demonstrate that Dora-VAE significantly outperforms existing methods across varying levels of shape complexity. Furthermore, we show that the improved reconstruction capability of Dora-VAE directly enhances the quality of downstream tasks by applying it to single-image 3D generation. The superior performance in generating geometric details validates our approach of focusing on salient region sampling and encoding for 3D VAE design.

Acknowledgements. This work is partially supported by the project L0751 between Bytedance and HKUST.

References

[1] Digital twin catalog. META, 2024. https://www.projectaria.com/datasets/dtc/. 5

[2] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 958–968, 2024. 1

[3] Yukang Cao, Liang Pan, Kai Han, Kwan-Yee K Wong, and Ziwei Liu. Avatargo: Zero-shot 4d human-object interaction generation and animation. *arXiv preprint arXiv:2410.07164*, 2024. 1

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 1

[5] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-δ: Fast and controllable image generation with latent consistency models, 2024. 1

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[7] Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1

[8] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174, 2016. 5

[9] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *Advances in Neural Information Processing Systems*, 35:30923–30936, 2022. 1

[10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 5

[11] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 5

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 5

[13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 5

[14] Fabian Groh, Patrick Wieschollek, and Hendrik P. A. Lensch. Flex-convolution (million-scale point-cloud learning beyond grid-worlds). In *Asian Conference on Computer Vision (ACCV)*, 2018. 2, 3

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1

[16] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[17] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d, 2024. 1

[18] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023. 7

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[20] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4), 2017. 7

[21] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *ECCV*, 2024. 1

[22] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3

[23] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *International Conference on Learning Representations (ICLR)*, 2024. 3

[24] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 1, 3, 7

[25] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 1

[26] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler,

Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[27] Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior, 2023. 3

[28] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. 2024. 1

[29] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, Hongzhi Wu, and Hao Su. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 3

[30] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 3

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[32] Carsten Moenning and Neil A Dodgson. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory, 2003. 3

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 8

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

[35] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9914–9925. IEEE, 2024. 3

[36] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4209–4219, 2024. 1, 3, 7

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[38] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 3

[39] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[40] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3

[41] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[42] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 1

[43] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 3

[44] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion, 2022. 1

[45] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[46] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 3

[47] Francis Williams, Jiahui Huang, Jonathan Swartz, Gergely Klar, Vijay Thakkar, Matthew Cong, Xuanchi Ren, Ruilong Li, Clement Fuji-Tsang, Sanja Fidler, Eftychios Sifakis, and Ken Museth. fvdb: A deep-learning framework for sparse, large-scale, and high-performance spatial intelligence. *ACM Transactions on Graphics (TOG)*, 43(4):133:1–133:15, 2024. 1

[48] Chengzhi Wu, Junwei Zheng, Julius Pfrommer, and Jürgen Beyerer. Attention-based point cloud edge sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[49] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv:2405.14832*, 2024. 1, 2, 3, 7

[50] Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang. OctFusion: Octree-based diffusion models for 3d shape generation. *arXiv*, 2024. 3

[51] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d

mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3

[52] Cem Yuksel. Sample elimination for generating poisson disk sample sets. *The Eurographs Association & John Wiley & Sons, Ltd.*, 2015. 3, 8

[53] Biao Zhang and Peter Wonka. Lagem: A large geometry model for 3d representation learning and diffusion. *arXiv preprint arXiv:2410.01295*, 2024. 3

[54] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 1, 3, 4, 5, 7, 8

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

[56] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 1, 2, 3, 5, 7, 8

[57] Xuying Zhang, Bo-Wen Yin, Yuming Chen, Zheng Lin, Yunheng Li, Qibin Hou, and Ming-Ming Cheng. Temo: Towards text-driven 3d stylization for multi-object meshes. *arXiv preprint arXiv:2312.04248*, 2023. 1

[58] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 3