# LATTICE: Democratize High-Fidelity 3D Generation at Scale

Zeqiang Lai[1,2⋆] , Yunfei Zhao[2⋆] , Zibo Zhao[2] , Haolin Liu[2]
Qingxiang Lin[2] , Jingwei Huang[2] , Chunchao Guo[2‡] , Xiangyu Yue[1‡]
[1]MMLab, CUHK [2]Tencent Hunyuan
https://lattice3d.github.io


Figure 1. High quality 3D assets generated by LATTICE from a single image.

## Abstract

*We present LATTICE, a new framework for high-fidelity 3D asset generation that bridges the quality and scalability gap between 3D and 2D generative models. While 2D image synthesis benefits from fixed spatial grids and well-established transformer architectures, 3D generation remains fundamentally more challenging due to the need to predict both spatial structure and detailed geometric surfaces from scratch. These challenges are exacerbated by the computational complexity of existing 3D representations and the lack of structured and scalable 3D asset encoding schemes. To address this, we propose VoxSet, a semi-structured representation that compresses 3D assets into a compact set of latent vectors anchored to a coarse voxel grid, enabling efficient and position-aware generation. VoxSet retains the simplicity and compression advantages of prior VecSet methods while introducing explicit structure into the latent space, allowing positional embeddings to guide generation and enabling strong token-level test-time scaling. Built upon this representation, LATTICE adopts a two-stage pipeline: first generating a sparse voxelized geometry anchor, then producing detailed geometry using a recitified flow transformer. Our method is simple at its core, but supports arbitrary resolution decoding, low-cost training, and flexible inference schemes, achieving state-of-the-art performance on various aspects, and offering a significant step toward scalable, high-quality 3D asset creation.*

⋆ Equal contribution. † Corresponding authors.

## 1. Introduction

Creating high-quality 3D assets is central to modern content pipelines across visual effects, gaming, virtual reality, and industrial design. Yet, manual creation remains labor-intensive and demands expert skills. Thus, automating 3D asset generation has become a key challenge at the intersection of vision, graphics, and machine learning.

Despite the impressive progress demonstrated by recent advances in 3D generation [46, 55, 57], the question of how to represent 3D assets remains *the "dark cloud" over scalable 3D generation* — a fundamental problem that continues to hinder progress in fidelity, efficiency, and generalization. This challenge is deeply intertwined *not only with classical 3D representations* — such as meshes, point clouds, Signed Distance Functions (SDFs), radiance fields [28], and 3D Gaussian Splattings [14] — *but also with VAE [6] representations* adopted by latent diffusion models [35], a key paradigm underpinning the recent advances in 3D generation, as well as in image and video synthesis [16, 17]. Even with latent diffusion–based compression, the underlying complexity of 3D structure and its cubic growth in memory and computation remain major obstacles, underscoring the need for compact representations. As a result, *compression, reconstruction, and generation* have become even more crucial in 3D learning, standing as long-standing themes throughout prior research.

The pursuit of such an ideal 3D representation has thus led to what we call *representation-centric* research, primarily revolving around compression and reconstruction by two leading paradigms: Sparse Voxel and VecSet. Sparse Voxel-based methods [34, 46] aim for efficiency by restricting computation to the active voxels near the object surface. However, as reported in Trellis [46], even with the inherent sparsity of 3D data, the sequence length of active voxels can be expensively long for training (over 20,000 at $64^3$ resolution), necessitating complex system designs based on sparse convolution [3] and attention mechanisms [4, 25], which leaves its scalability an open question. Nonetheless, the structured latent space provides strong flexibility for editing and broad generalization to diverse downstream tasks [5, 12, 19, 51]. VecSet-based approaches [54, 55, 57] offer a more compact and elegant alternative by compressing 3D objects into a small set of feature vectors via cross-attention between densely sampled point cloud and sparsely sampled point queries – a set of point coordinates uniformly sampled in the object surface. Remarkably, as few as 3,072 vectors can already yield excellent reconstruction quality, making VecSet-based models highly efficient. Moreover, all operations within these models — including the VAE and DiT components — can be implemented using standard self- and cross-attention layers, enabling excellent scalability within modern transformer architectures. Despite the strong advances in compression and reconstruction, current models remain notably behind 2D latent diffusion models in quality and scalability, leaving the enhancement of 3D generation capability an open and underexplored challenge.

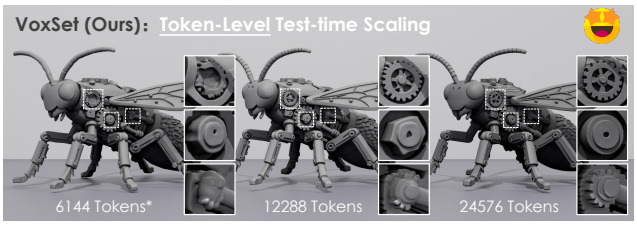


Figure 2. Illustration of test-time scaling in our model. The model is trained with up to 6,144 tokens, but is evaluated under different token counts at test time, showing notable improvements.

In this paper, we seek to answer a central question in a *generation-centric* perspective, *i.e.*,

> What truly defines a good representation for 3D diffusion **generator** itself?

From this viewpoint, we start by asking – *why does 3D generation still significantly lag behind 2D in quality and scalability?* At its core, the gap stems from a fundamental difference in how the generative task is framed. In 2D image synthesis, the spatial grid is predefined – models only need to infer RGB values at fixed pixel coordinates (a secret condition that greatly simplifies the denoising processes). 3D generation, however, faces a far more open-ended task: it must discover both **where** to place content in space and **what** to represent there (*e.g.*, SDF, RGB). This joint reasoning over structure and content dramatically expands the search space and introduces ambiguity, making optimization harder and scaling behavior less predictable[1].

At first glance, Sparse Voxel may seem a promising choice due to its inherent spatial structure. While this holds true, we instead still prefer a *VecSet*-based representation for its distinctive advantages we detailed later, more importantly, its strong capability for *test-time scaling*, as illustrated in Fig. 2. Building upon the insights discussed earlier, we demonstrate that it is possible to combine the best of both worlds through *Localizable Code* – a unified and high-level abstraction for any representation that tackles the joint reasoning problem. Crucially, it is localizability rather than structure that truly matters. Guided by this principle, our key idea is to add localizable structure to VecSet [54] latent codes. In other words, we aim to decouple the prediction of *where* and *what*, and guide the unstructured VecSet generation with structure/position, mimicking the success of image generation based on a 2D grid.

To achieve this, we investigate the positional information secretly encoded in VecSet latent produced by point queries, *i.e.*, each latent is strongly correlated with regions

[1] See Appendix for more discussions.

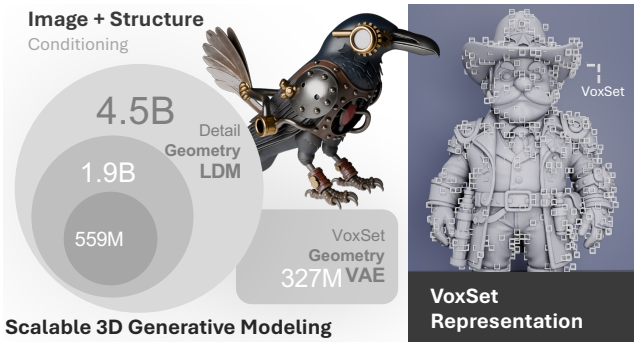


Figure 3. LATTICE system: At its core is a novel VoxSet representation, enabling scalable 3D modeling from 0.6B to 4.5B.

near the position of its corresponding point query, as hinted in [18]. However, this information can hardly help during the generation as the positions of point queries are unknown at test time. Therefore, we introduce *voxel queries* as a perfect replacement of point queries. Instead of utilizing point coordinates on the surface, we use the center coordinates of the active voxel intersecting the object's surface. These active voxel grids can be very coarse, thus can be easily obtained, during the test time, by voxelizing an existing geometry generated by any off-the-shelf geometry generation models [13, 46] with less perfect quality.

As a result, our first improvement leads to *VoxSet*, a new semi-structured representation that inherits the efficiency and simplicity of VecSet [54], while introducing structure into its latent space. This design brings several unique benefits: (1) it enables *flexible encoding and decoding in abritrary resolution*, which makes training particularly cheap as multi-stage training is possible by pretraining on very low token size and progressively scaling the tokens up; (2) every latent in *VoxSet is structured, anchored in a 3D regular voxel grid* so that the position information can be directly injected into the diffusion transformer (DiT) [31] through positional embedding [40, 42], which provides strong guidance during diffusion generation and is proven to be essential in model scaling in our experiment.

Based on VoxSet, we present *LATTICE*, a general framework designed to generate high-fidelity and detailed 3D assets. LATTICE employs a two-stage pipeline. In the first stage, it generates a sparse voxel grid by voxelizing a coarse mesh produced by any off-the-shelf model, such as Hunyuan3D-2 [57] or Trellis [46]. In the second stage, it generates geometry VoxSets at arbitrary resolutions (number of tokens) within the selected voxel grid. Built on rectified flow transformer [17] and a progressive training strategy, we train a family of large-scale image-to-3D generation models — with up to 4.5 billion parameters, as shown in Fig.3, capable of producing detailed meshes from a single image. Through extensive evaluation, we demonstrate that LATTICE exhibits strong superiority against previous state-of-the-art models, and is distinguished by several key strengths, as summarized below:

- **Test-time scaling.** Our model exhibits a strong test-time scaling effect. The model trained with up to 6144 tokens/voxel cells can be directly scaled to up to 30720 tokens during the test time, with consistent improvement.
- **Low-cost training.** Our base model, with 2 billion parameters, can be effectively trained in under 24 hours using 64 GPUs, while still significantly outperforming previous methods.
- **Simplicity.** The model architecture is exceptionally simple — relying solely on a pure Transformer design, without any complex or sparse components.
- **Exceptional performance.** Our model achieves significantly strong performance in 3D generation, excelling in geometry smoothness and detail preservation.

LATTICE represents a significant step forward in next-generation 3D assets generation, bridging the gap between generated and handcrafted 3D assets. We hope this work offers valuable insights into effective scaling of 3D generation models and opens up new possibilities for automated, high-fidelity 3D content creation.

## 2. Related Works

### 2.1. 3D Representations

Unlike images and videos, which are universally represented by pixel colors, 3D assets exhibit a wide variety of representations tailored to different application contexts [24, 28, 38, 47]. Common atomic representations include voxels, point clouds, Signed Distance Fields (SDF), polygon meshes, DMTet [37], Flexicube [38], Neural Radiance Fields (NeRF) [28], and Gaussian Splatting [14], among others. These representations, whether explicit or implicit, serve distinct roles in the 3D industry—for example, point clouds are prevalent in perception tasks like autonomous driving [33], NeRF excels in novel view rendering, and polygon meshes remain the standard for gaming and real-time applications. These atomic representations can often be converted between one another—for example, polygon meshes can be extracted from SDFs via the marching cubes algorithm [26]. Ultimately, the choice of representation is task-dependent and directly influences network design, *e.g.*, autoregressive models [43] for meshes, and diffusion models [58] for SDF.

Nonetheless, even lightweight and flexible representations such as implicit functions still impose significant modeling and computational burdens on deep neural networks, especially diffusion models [35] – the current golden paradigm for 3D generation. As a result, the latent representations of 3D assets have emerged as a new research focus, aiming to enhance efficiency. These representations,

with difference by their own, can be generally categeoried into three popular types, *i.e.*, (1) VecSet, represented by 3DShape2VecSet [54], compress 3D shapes into 1D latent sets; (2) Triplane, represented by Direct3D [11], compress shapes into three orthogonal features planes; and (3) Sparse Voxel, represented by XCube [34], converted 3D assets into features anchored on sparse voxels. A long-standing belief holds that the spatial locality of sparse voxel representations helps preserve fine details, whereas VecSet representations, despite their efficiency, tend to lose details due to their global modeling. In this paper, we challenge this idea and identify that the key for 3D generative models lies not in locality, but in a well-known structure at test time. Here, we introduce VoxSet, a semi-structured latent representation that combines efficiency and strong expressiveness.

### 2.2. Geometry Generation

3D geometry generation has advanced rapidly in recent years. Early works [36, 44, 48, 53] based on different generative models [7, 15, 30] demonstrated the preliminary potential for generating specific categories of geometry. With the rise of diffusion models [10, 35], 3D geometry generation methods based on score distillation [32] have been introduced, enabling text-to-3D generation by leveraging text-to-image models. Feedforward methods such as LRM [11], Hunyuan3D 1.0 [50], and LGM [41] represent another line of research focused on generating 3D assets in a single step. On the other hand, autoregressive models, *e.g.* MeshGPT [39], BPT [43], and Meshtron [8] have become popular for mesh generation with human-like topology.

Recently, native 3D diffusion models have significantly improved generation quality by utilizing 3D data. Notable examples include Michelangelo [56], CLAY [55], Hunyuan3D 2.0 [57], TripoSG [22], and Step1X-3D [21], building on 3DShape2VecSet [54]. Despite great success, these methods seem to struggle at generating highly detailed meshes. On the contray, another line of research, following XCube [34], shows promising results by works as Trellis [46], Hi3DGen [52], SparseFlex [9], Sparc3D [23] and Direct3D-s2 [45]. Nevertheless, we show that it is effective scaling through localizable guidance, rather than VecSet or XCube, that matters in detailed geometry generation.

## 3. Scalable 3D Generative Modeling

The goal of LATTICE is to explore a new paradigm for scalable 3D generative modeling. To achieve this, a key design choice in our approach is utilizing a coarse geometry structure as strong guidance for detailed geometry generation. This design, despite being used in many voxel-based approaches, such as XCube [34] and SLAT [46], is significantly underestimated. In this work, we show that it is actually essential for effective model scaling and performance improvement, no matter what the representation is.

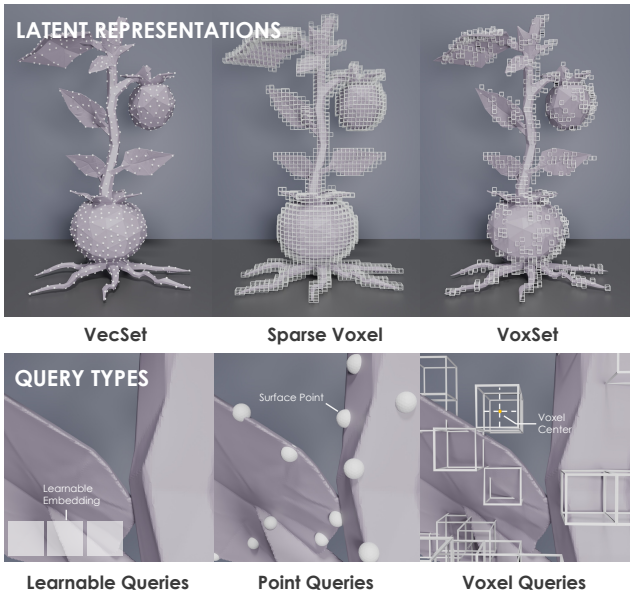


Figure 4. Illustrations of different latent representations and different query types.

### 3.1. VoxSet Representation

Underlying the architecture of LATTICE, it is *VoxSet* representation that builds up the core of the entire system. Existing 3D representations—such as meshes, point clouds, signed distance field (SDF), NeRF [28], FlexiCubes [38], VecSet [54], and SLAT [46]—can be broadly categorized into two types. The first includes explicit or implicit atomic representations, such as point clouds, SDFs, and NeRF, which directly encode geometry or appearance. The second includes latent representations, such as VecSet and SLAT, which are built upon atomic representations with variational autoencoders (VAE) [6, 15] and tailored for building compact latent space in latent diffusion models [35].

VoxSet is a latent representation guided by two key principles: scalability and structural latent space. To support scalability, VoxSet compresses any 3D asset into a sequence of latent tokens via a cross-attention mechanism, following the design of 3DShape2VecSet [54]. Formally, given a 3D object, we employ a VAE to encode its point cloud representation and reconstruct the corresponding SDF, from which a surface mesh can be extracted via the Marching Cubes algorithm [26]. The input point cloud $P \in \mathbb{R}^{N\times 7}$ captures multiple attributes per point, where $N$ denotes the total number of points and each point encodes its 3D coordinates, surface normal, and a binary sharpness indicator marking whether it lies on a sharp edge. Following the strategy in Hunyuan3D-2 [57], the point cloud is constructed by combining uniform sampling over the surface with importance sampling around sharp edges to better preserve high-frequency details.

**Efficient Scaling via Sparsity.** The latent represen-

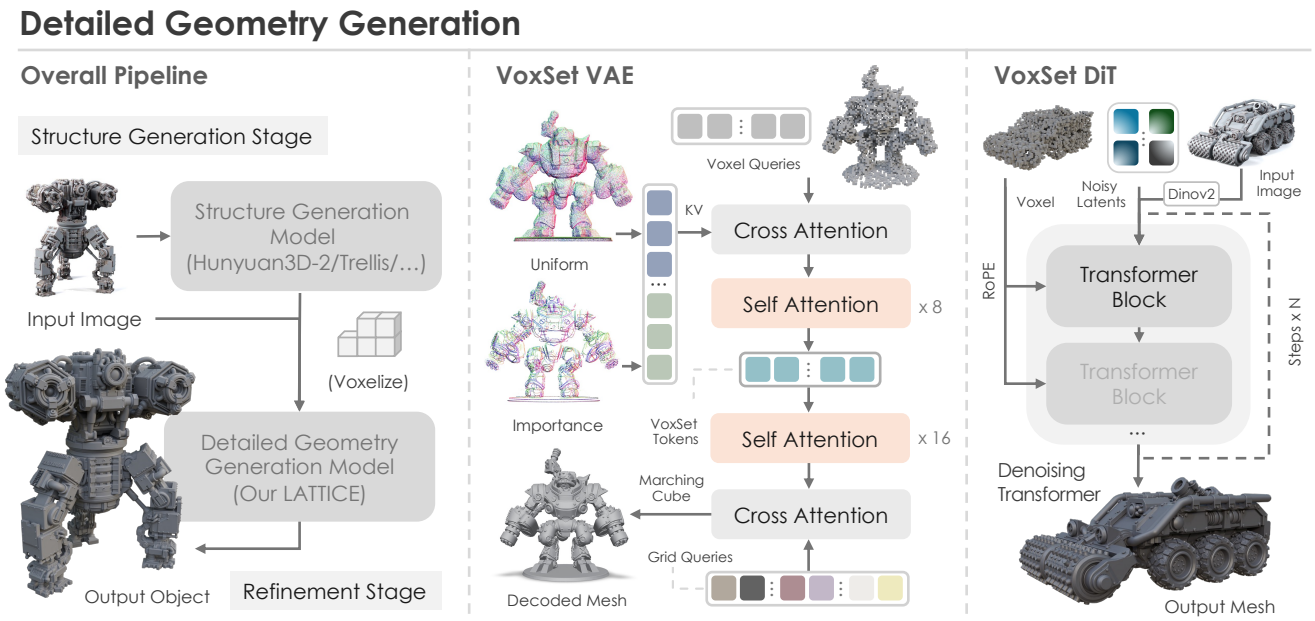


Figure 5. **LATTICE Model Architecture**: it features a two-stage coarse-to-fine pipeline and a novel VoxSet VAE and DiT.

tation, *i.e.*, a token sequence, is obtained via performing cross-attention between the input point cloud and a set of query tokens following a series of self-attention layers. The decoder is designed symmetrically, where the SDF grid coordinates serve as queries of the cross-attention against latent tokens. Notably, these latent tokens, in fact, secretly encode the global signals; thus we could represent any 3D object with a latent sequence of any length [2, 55, 57]. This is particularly useful as progressive token scaling (a more fine-grained strategy than progressive resolution scaling) is possible to greatly reduce the training cost, *e.g.*, starting pretraining from 1024 tokens and progressively increasing to more. Even more, as evidenced in FlashVDM [18], a diffusion model trained on 512 latent tokens can be directly scaled up to 3072 tokens **in test-time** with better performance, which makes VoxSet particularly economical.

**Voxel Queries for Detail Modeling.** The choice of query set is a crucial design choice in VecSet-like methods [54, 56], which also serves as a key distinguishing factor of VoxSet. In 3DShape2VecSet [54], two types of query set, *i.e.*, learnable queries and point queries, are proposed. The learnable queries encode the global statistics and are easy to train, but are limited in scaling up for better reconstruction and generation performance. Point queries are downsampled point clouds with furthest point sampling, which encode local information around the queries and support encoding-decoding at arbitrary resolution, favoring low-cost progressive scale-up. Moreover, the locality of latent tokens encoded by point queries is very strong and correlated with their position, as discussed in FlashVDM [18].

In other words, the latent set from point queries is ordered with position information secretly encoded. However, none of existing models [22, 55, 57] utilize the information. One of the biggest obstacles is that point queries are sampled on the object surface, whose positions are unknown during test time. To address this problem, we introduce *Voxel Queries*, a query set anchored at the center of active voxels intersecting with the object surface, as shown in Fig. 4. Voxel Queries are not sampled on the surface but on a coarse voxel grid, so that their position can be easily obtained during test time by a coarse structure generation stage. Besides, the voxel center is decorrelated with different surfaces, reducing the training-test gap and greatly improving the generalization capabilities at test time.

### 3.2. Detailed Geometry Generation

As shown in Fig. 5, we introduce a two-stage pipeline to fit the proposed VoxSet representation for generating geometry with ultimate details. The first stage generates the coarse sparse structure given the input image by voxelizing the results of off-the-shelf pretrained 3D generators [57]. The second stage generates sparse voxel latents anchored at the voxel centers of the previous sparse structure.

**Semi-Structured Geometry VAE.** We adopt the proposed VoxSet representation to train a semi-structured geometry VAE. As illustrated in Fig. 4, our method combines the strengths of VecSet [54] and SLAT [46], keeping latent tokens compact and structural. To support multi-resolution voxel structures, instead of randomly sampling voxel queries at various resolutions, we propose a simpler

approach that supports arbitrary resolutions. Specifically, we jitter the point queries during training by adding a small random offset $\epsilon \sim U\left[\frac{-1}{2R}, \frac{1}{2R}\right]$, where $R$ is the smallest resolution we aim to support. At test time or during diffusion training, voxel queries can be sampled at any resolution greater than $R$. The other aspects of our VAE are the same as in Hunyuan3D-2 [57], except we only sample queries from a uniformly sampled point cloud.

**Adding Structure to Diffusion Transformer.** Following Hunyuan3D-2 [57], we utilize a rectified-flow transformer to generate the VoxSet. Instead of solely conditioning on the input image, we propose to utilize the structure of VoxSet by adding rotary positional embedding (RoPE) [40] to each noisy latent token. This change, despite being inconspicuous at first glance, is crucial in improving model convergence. The reason behind this can be two-fold, firstly, the amount of available 3D data is much smaller than 2D image and video counterparts, which makes the latent space severely unoccupied. Secondly, geometry generation is drastically different and a more difficult task than image generation due to its sparsity, *i.e.*, the 3D geometric surface occupies only a small portion of its bounding box while every pixel in the image has an RGB value. As a result, previous approaches [55, 57] that only use a single image as condition could hardly guide the denoising trajectory towards detailed geometry. To reduce the training cost, we introduce two simple strategies: (1) instead of utilizing all structure tokens, we randomly sampled a fixed number of tokens during the training, which is much smaller than sparse voxel methods [46]; (2) we adopt a progressive training strategy by first training on 1024 tokens and progressively scaling up to 6144 tokens.

**Image Conditioning.** Following Hunyuan3D-2 [57], we use Dinov2-Giant [29] for image conditioning, taking the last hidden layer embedding without the class token. Different from Hunyuan3D-2's 518 resolution, we use 1022 for finer details. The object is cropped via a binary mask while keeping the aspect ratio to reduce token length. No extra positional embedding is added, as Dino already encodes sufficient spatial information.

**Training and Test-time Scaling.** We train several models in different model sizes, ranging from 0.6B to 4.5B. As shown in Fig. 6, our model exhibits stable scaling effect – the bigger the better. Moreover, our model surprisingly reveals test-time scaling effect in token length as shown in Fig. 6. Even though our model is trained on 6144 tokens, we could increase the number of tokens to 12288, 24576, and even more by sampling more voxel queries.

### 3.3. Applications

Thanks to the flexible design of the proposed architecture. We could adapt our model for various tasks.

**Mesh Refinement** can be extended to a broader context where the image is not aligned or missing, such as controllable generation or part refinement [49] as in Fig.7.

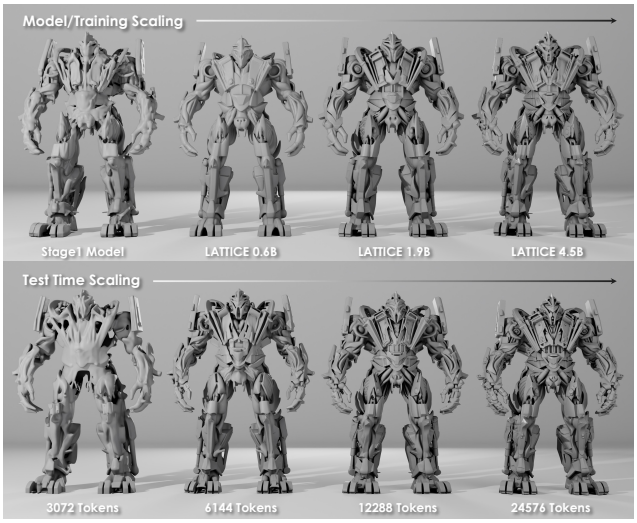


Figure 6. Illustration of model/training and test scaling effects.

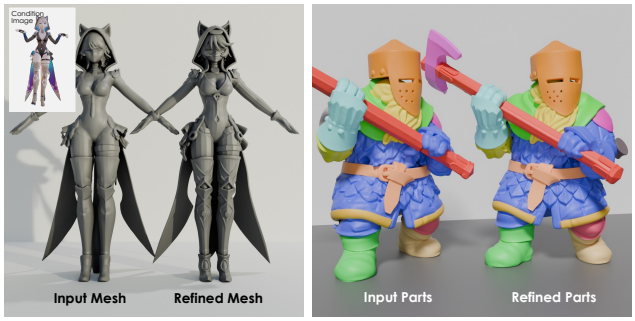


Figure 7. Illustrations of applications of LATTICE. Mesh refinement in the left and part refinement in the right.

**Mesh Editing** is also possible by manipulating the voxel queries and latent features of the given mesh, using the idea of Repaint [27], MastControl [1], *etc*.

## 4. Experiments

### 4.1. Reconstruction

We adopt two metrics for evaluating geometry reconstruction performance, including Chamfer Distance (CD) and F-score with a threshold of 0.001. To evaluate the reconstruction accurately, we use points-to-surface distances to calculate the metrics, with the mesh normalized to the range [-1, 1]. Similar to Dora [2], we construct a benchmark containing more challenging, detailed assets as LATTICE-Bench(R). The competing methods consist of (1) representative VecSet-based methods: Hunyuan3D-2 [57]; and (2) Voxel-based methods: SparseFlex [9], and Direct3D-s2 [45]. The numerical comparison is shown in Tab. 1, the metrics are multiplied by $10^4$ and $10^2$. Our method delivers


Figure 8. Visual comparison of geometry generation against several state-of-the-art open-source methods.

top performance with a much more compact latent representation than voxel-based methods.

### 4.2. Generation.

We evaluate the image-to-geometry generation through various metrics including ULIP [47], Uni3D [59] for text-mesh and image-mesh similarities, following Hunyuan3D-2 [57]. We compare our method against (1) open-source methods, Michelangelo [56], Craftman 1.5 [20], Trellis [46], Hunyuan3D-2 [57], Hi3DGen [52], and Direct3D-s2 [45]; (2) closed-source methods, which we denote as Model 1-4. The numerical comparison is shown in Tab. 2, omitting closed-source methods as obtaining a large amount of their results is very expensive. We compare LATTICE-1.9B, which is the closest size to other models. It can be observed that our method achieves the best performance. Fig. 8 and Fig. 9 demonstrates the visual comparison, which confirms the superiority of our method.

### 4.3. Evaluation

**Effect of Voxel Queries.** To assess the effectiveness of the proposed voxel queries, we compare three DiTs, each uti-

| Method | Res- | Latent Size | CD(↓) | F1(↑) |
|---|---|---|---|---|
| Hunyuan3D-2 [57] | N/A | 64 × 4096 | 12.35 | 82.78 |
| | N/A | 64 × 8192 | 9.157 | 91.57 |
| SparseFlex [9] | 512 | 8 × 48557 | 8.020 | 90.94 |
| | 1024 | 8 × 196028 | 2.972 | 97.76 |
| Direct3D-s2 [45] | 1024 | 64 × 46592 | 4.987 | 97.46 |
| **LATTICE** (Ours) | N/A | 64 × 4096 | 5.321 | 95.31 |
| | N/A | 64 × 8192 | 2.909 | 98.53 |
| | N/A | 64 × 20480 | **1.893** | **99.59** |

Table 1. Quantitative comparisons of geometry reconstruction.

| Method | ULIP-T | ULIP-I | Uni-T | Uni-I |
|---|---|---|---|---|
| Michelangelo [56] | 0.075 | 0.115 | 0.213 | 0.261 |
| Craftsman 1.5 [20] | 0.074 | 0.129 | 0.237 | 0.298 |
| Trellis [46] | 0.076 | 0.126 | 0.249 | 0.311 |
| Hunyuan3D 2.0 [57] | 0.077 | 0.130 | 0.251 | 0.315 |
| Hi3DGen [52] | 0.066 | 0.112 | 0.246 | 0.299 |
| Direct3D-s2 [45] | 0.074 | 0.122 | 0.247 | 0.314 |
| **LATTICE**-1.9B | **0.078** | **0.130** | **0.254** | **0.315** |

Table 2. Numerical comparison of geometry generation performance on ULIP [47] and Uni3D [59] similarities.


Figure 9. Visual comparison against commercial models.


Figure 10. Ablation study on the proposed voxel query and VoxSet VAE, by incrementally adding each component.

lizing different VAEs and query types, as shown in Fig. 10. All DiTs were trained for 200k steps (100k on 1024 tokens and 100k on 3072 tokens). The results indicate that voxel queries produce fewer artifacts, benefiting from a reduced domain gap, and the VoxSet VAE introduces more detail

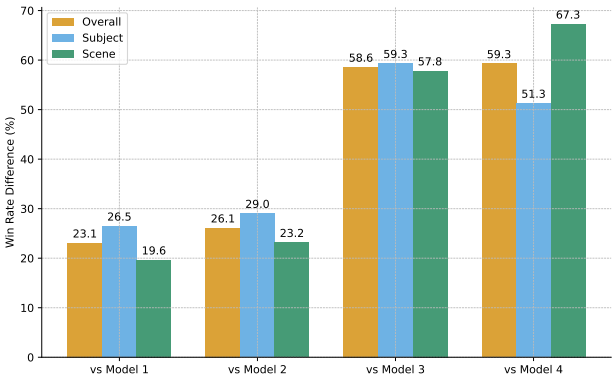


Figure 11. User study of our method against competitors showing win rate (%) across Overall, Subject, and Scene categories.

| | Baseline | | + Fixed Train | | | + Query Jitter | | |
|---|---|---|---|---|---|---|---|---|
| **Res-** | 64 | 128 | 64 | 128 | 256 | 64 | 128 | 256 |
| CD(↓) | 10.7 | 7.72 | 6.42 | 5.73 | 5.69 | 6.03 | 5.32 | 5.36 |
| F1(↑) | 85.3 | 91.4 | 92.9 | 94.5 | 94.7 | 93.7 | 95.3 | 95.3 |

Table 3. Ablation study of VAE training strategies. All settings are tested with 4096 tokens and voxel queries.

thanks to better reconstruction capability.

**Effect of Query Jitter.** Voxel queries are essential for bridging the gap between training and testing results in the first stage. To evaluate their impact, we ablate several VAEs with voxel queries by assessing their reconstruction performance. The numerical comparison is presented in Tab. 3. As shown, the original point-query VAE suffers a significant degradation when tested with voxel queries. In contrast, the Query Jitter VAE outperforms VAEs trained at a fixed resolution and offers greater flexibility when applied to varying resolutions.

**User Study.** We also conducted a user study to assess human preferences across different methods. As shown in Fig. 11, our method was compared to four commercial models. The results clearly show that our method significantly outperforms the others.

## 5. Conclusion

We have presented LATTICE, a novel framework that advances 3D asset generation by introducing Voxset, a semi-structured latent representation. By conditioning on localizable position information, we address key challenges in computational complexity, scalability, and fidelity for diffusion generation. Our method demonstrates superior performance in generating high-quality meshes, achieving stunning detail, smoothness, and sharpness. With its flexible encoding, low-cost training, and strong test-time scaling, LATTICE represents a significant step forward in the automated generation of scalable, high-fidelity 3D content.

## Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (No. 62306261), HK RGC-Early Career Scheme (No. 24211525), ITSP Platform Project (No. ITS/600/24FP) and the SHIAE Grant (No. 8115074). This study was supported in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This work is also partially supported by Hong Kong RGC Strategic Topics Grant (No. STG1/E-403/24-N), and CUHK-CUHK(SZ)-GDST Joint Collaboration Fund (No. YSP26-4760949).

## References

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 6

[2] Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. *arXiv preprint arXiv:2412.17808*, 2024. 5, 6

[3] Spconv Contributors. Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv, 2022. 2

[4] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2024. 2

[5] Paul Engstler, Aleksandar Shtedritski, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Syncity: Training-free generation of 3d worlds. *arXiv preprint arXiv:2503.16420*, 2025. 2

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 4

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[8] Zekun Hao, David W Romero, Tsung-Yi Lin, and Ming-Yu Liu. Meshtron: High-fidelity, artist-like 3d mesh generation at scale. *arXiv preprint arXiv:2412.09548*, 2024. 4

[9] Xianglong He, Zi-Xin Zou, Chia-Hao Chen, Yuan-Chen Guo, Ding Liang, Chun Yuan, Wanli Ouyang, Yan-Pei Cao, and Yangguang Li. Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. *arXiv preprint arXiv:2503.21732*, 2025. 4, 6, 8

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 4

[11] 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. *arXiv preprint arXiv:2311.04400*, 2023. 4

[12] Binbin Huang, Haobin Duan, Yiqun Zhao, Zibo Zhao, Yi Ma, and Shenghua Gao. Cupid: Pose-grounded generative 3d reconstruction from a single image. *arXiv preprint arXiv:2510.20776*, 2025. 2

[13] Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025. 3

[14] 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. 2, 3

[15] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3

[18] Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Haolin Liu, Fuyun Wang, Huiwen Shi, Xianghui Yang, Qinxiang Lin, Jinwei Huang, Yuhong Liu, Jie Jiang, Chunchao Guo, and Xiangyu Yue. Unleashing vecset diffusion model for fast shape generation, 2025. 3, 5

[19] Lin Li, Zehuan Huang, Haoran Feng, Gengxiong Zhuang, Rui Chen, Chunchao Guo, and Lu Sheng. Voxhammer: Training-free precise and coherent 3d editing in native 3d space. *arXiv preprint arXiv:2508.19247*, 2025. 2

[20] Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner, 2024. 7, 8

[21] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 4

[22] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025. 4, 5

[23] Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025. 4

[24] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding, 2023. 3

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[26] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 3, 4

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and L Repaint Van Gool. Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2023. 6

[28] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 4

[29] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 6

[30] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. 4

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 3

[32] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 4

[33] 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. 3

[34] 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. 2, 4

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4

[36] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 4

[37] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[38] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 3, 4

[39] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19615–19625, 2024. 4

[40] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3, 6

[41] 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. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 4

[42] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[43] Haohan Weng, Zibo Zhao, Biwen Lei, Xianghui Yang, Jian Liu, Zeqiang Lai, Zhuo Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, et al. Scaling mesh generation via compressive tokenization. *arXiv preprint arXiv:2411.07025*, 2024. 3, 4

[44] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 4

[45] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Philip Torr, Xun Cao, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. 4, 6, 7, 8

[46] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2, 3, 4, 5, 6, 7, 8

[47] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 3, 7, 8

[48] 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. 4

[49] Xinhao Yan, Jiachen Xu, Yang Li, Changfeng Ma, Yunhan Yang, Chunshi Wang, Zibo Zhao, Zeqiang Lai, Yunfei Zhao, Zhuo Chen, et al. X-part: high fidelity and structure coherent shape decomposition. *arXiv preprint arXiv:2509.08643*, 2025. 6

[50] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d-1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024. 4

[51] Yunhan Yang, Yufan Zhou, Yuan-Chen Guo, Zi-Xin Zou, Yukun Huang, Ying-Tian Liu, Hao Xu, Ding Liang, Yan-Pei Cao, and Xihui Liu. Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. *arXiv preprint arXiv:2507.06165*, 2025. 2

[52] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 3:2, 2025. 4, 7, 8

[53] Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Jiayuan Fan, Gang Yu, Taihao Li, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. *arXiv preprint arXiv:2311.17618*, 2023. 4

[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 (TOG)*, 42(4):1–16, 2023. 2, 3, 4, 5

[55] 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. 2, 4, 5, 6

[56] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 5, 7, 8

[57] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 2, 3, 4, 5, 6, 7, 8

[58] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4):1–13, 2023. 3

[59] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 7, 8