TAR3D: Creating High-Quality 3D Assets via Next-Part Prediction

Xuying Zhang[1][*][§], Yutong Liu[3][*], Yangguang Li[4,5], Renrui Zhang[4], Yufei Liu[6], Kai Wang[1],
Wanli Ouyang[4,6], Zhiwei Xiong[3], Peng Gao[6], Qibin Hou[1,2][†], Ming-Ming Cheng[1,2]

[1]VCIP, CS, Nankai University   [2]NKIARI, Shenzhen Futian   [3]USTC

[4]CUHK MMLab   [5]VAST   [6]Shanghai AI Lab

Homepage: https://github.com/HVision-NKU/TAR3D

Figure 1. 3D asset gallery generated by the proposed TAR3D model. We employ SyncMVD [42] as the texture synthesizer.

Abstract

We present TAR3D, a novel framework that consists of a 3D-aware Vector Quantized-Variational AutoEncoder (VQ-VAE) and a Generative Pre-trained Transformer (GPT) to generate high-quality 3D assets. The core insight of this work is to migrate the multimodal unification and promising learning capabilities of the next-token prediction paradigm to conditional 3D object generation. To achieve this, the 3D VQ-VAE first encodes a wide range of 3D shapes into a compact triplane latent space and utilizes a set of discrete representations from a trainable codebook to reconstruct fine-grained geometries under the supervision of query point occupancy. Then, the 3D GPT, equipped with a custom triplane position embedding called TriPE, predicts the codebook index sequence with prefilling prompt tokens in an autoregressive manner so that the composition of 3D geometries can be modeled part by part. Extensive experiments on several public 3D datasets demonstrate that TAR3D can achieve superior generation quality over existing methods in text-to-3D and image-to-3D tasks.

---

[*]Equal contribution.   [†]Corresponding author.
[§]Work done during the internship at Shanghai AI Lab.

1. Introduction

Conditional 3D object generation aims to generate high-quality 3D assets that semantically conform to the given prompts, *e.g.,* 2D images, and text. It has witnessed significant progress with substantial advances in diffusion-based methods, including Score Distillation Sampling (SDS) optimization [7, 49], multiview synthesis [40, 55, 73], and 3D-aware diffusion generation [33, 72, 80]. However, establishing a unified model across various modalities poses huge challenges due to the different paradigms between diffusion models and large language models (LLMs) [1, 3, 13, 45].

More recently, pioneering works like MeshGPT [56] and MeshXL [8] attempt to introduce the autoregressive manner of LLM into 3D mesh generation. However, they struggle in a sequence length of 9 times the face number, which is attributed to their quantization of the vertices of mesh faces. Such an excessively long sequence length limits their applicability in industrial 3D assets with hundreds of thousands of faces. This motivates us to investigate whether it is possible to encode a mesh with an arbitrary number of faces into a fixed-length sequence to alleviate the computational burden of autoregressive modeling.

To address this question, the key is to efficiently encode the 3D representations such that the sequence length is not proportional to the number of faces. Triplane representation provides such a good property. Different from other 3D representations like unordered point clouds or verbose voxels, this representation is able to balance storage efficiency with strong expressiveness by compressing 3D information into three 2D feature maps with a fixed size. Due to the high-level 2D compositions, it fits discretization better and can also take advantage of the strategies from image quantization methods, e.g., LLamagen [59]. Based on triplane representation, we propose TAR3D, a novel 3D autoregressive framework consisting of a 3D VQ-VAE and a 3D GPT.

The 3D VQ-VAE encodes the 3D shapes into compact triplane features, from which a trainable codebook of context-rich 3D geometry parts is employed to acquire a set of discrete embeddings. In this way, the 3D meshes can be represented as a feature sequence with triplane-size length, regardless of the number of faces, reducing the reliance on huge GPU resources. The quantized latent representations are then decoded to a neural occupancy field for 3D reconstruction. To achieve information exchange among different planes, we propose incorporating feature deformation and attention mechanism designs during the decoding process for fine-grained geometry details. The 3D GPT is driven by prefilling prompt embeddings to model the codebook index sequence corresponding to quantized triplane features, enabling conditional 3D object generation in an autoregressive manner. To preserve more spatial information, during generation, we also custom-craft a 3D positional encoding strategy dubbed TriPE, in which the 2D positional information of each plane and the 1D ones between the same position of the three planes are organically fused.

To validate the effectiveness of our TAR3D, we conduct extensive experiments on a wide range of 3D objects of two popular benchmark datasets, i.e., ShapeNet [6] and Objaverse [17], and an out-of-domain dataset, Google Scanned Objects [21]. Based on the autoregressive manner and our well-designed strategies,TAR3D can create high-quality 3D assets, as shown in Fig. 1. The quantitative and qualitative results also demonstrate that our TAR3D can significantly outperform recent cutting-edge 3D generation methods.

Our contributions can be summarized as follows:

- We present TAR3D, a novel autoregressive pipeline composed of a 3D VAE and a 3D GPT for conditional 3D object generation. To our knowledge, this is the first attempt to quantize the 3D objects with triplane representations and generate high-quality assets part by part.
- We introduce feature deformation and attention mechanism designs to capture fine-grained geometry details.
- We propose a 3D position encoding strategy, i.e., TriPE, to preserve the spatial information as much as possible.

2. Related Work

2.1. 3D Generation

Early 3D generation approaches mainly focus on the generation of different forms of 3D models, e.g., point clouds [30, 71, 82], meshes [9, 74, 87, 88], and volumes [5, 58, 66] in a text/image-conditioned or unconditional manner. Limited by the categories and quantity of 3D objects used for training, these methods often do not generalize well. With the remarkable progress achieved by diffusion models in 2D image generation, a large number of researchers have explored migrating the pre-trained 2D priors to 3D generation. Pioneer works [7, 49, 65, 84] feed the rendered views to a 2D pre-trained model and perform per-shape optimization for knowledge distillation. Despite ushering in a new era, these methods suffer from a series of serious issues, e.g., time-consuming and multi-face. Another line of methods like Zero123/++ [40, 55], One2345 [38], and AR123 [85] use the rendered views of 3D objects to finetune the pre-trained diffusion models for new-view or multi-view generation. To improve the consistency between the generated multiple views, Consistent123 [37] and Cascade-Zero123 [10] introduce extra priors, e.g., boundary, and redundant views. SyncDreamer [41] proposes to correlate the corresponding features across different views by building a 3D-aware attention mechanism. These methods can be followed by a sparse-view reconstruction model, e.g., NeRFs [46, 48, 67], LRMs [25, 27, 73] and LGM [62], to generate 3D objects. However, the indirect generation fashion of these methods may lead to detail loss or reconstruction failures due to their heavy reliance on the fidelity of multi-view images.

More recently, a surge of works [14, 72, 78, 86] directly synthesize 3D objects via a 3D diffusion model, in which the 3D shapes are fed to a pre-trained 3D VAE for continuous latent features. Unlike these methods, our TAR3D eschews the diffusion scheme and creates 3D objects by autoregressively generating discrete geometric parts.

2.2. VAE & VQ-VAE

Variational Autoencoder (VAE) [32] is usually utilized to map high-dimensional input information to continuous probabilistic latent representations and has a far-reaching impact in the field of generative modeling. Thanks to this work, recent diffusion models [12, 53, 54, 60] can be trained on limited computational resources while retaining their quality and flexibility. Recent 3D generation approaches like Clay [80], Direct3D [72], and LN3Diff [33] also explore to construct 3D-aware diffusion models based on this technology. Vector Quantised-Variational AutoEncoder (VQ-VAE) [64] is a variant of VAE. It introduces the codebook mechanism to quantize the continuous latent representations into discrete components, achieving promising performance on many generative tasks, e.g., text-to-image
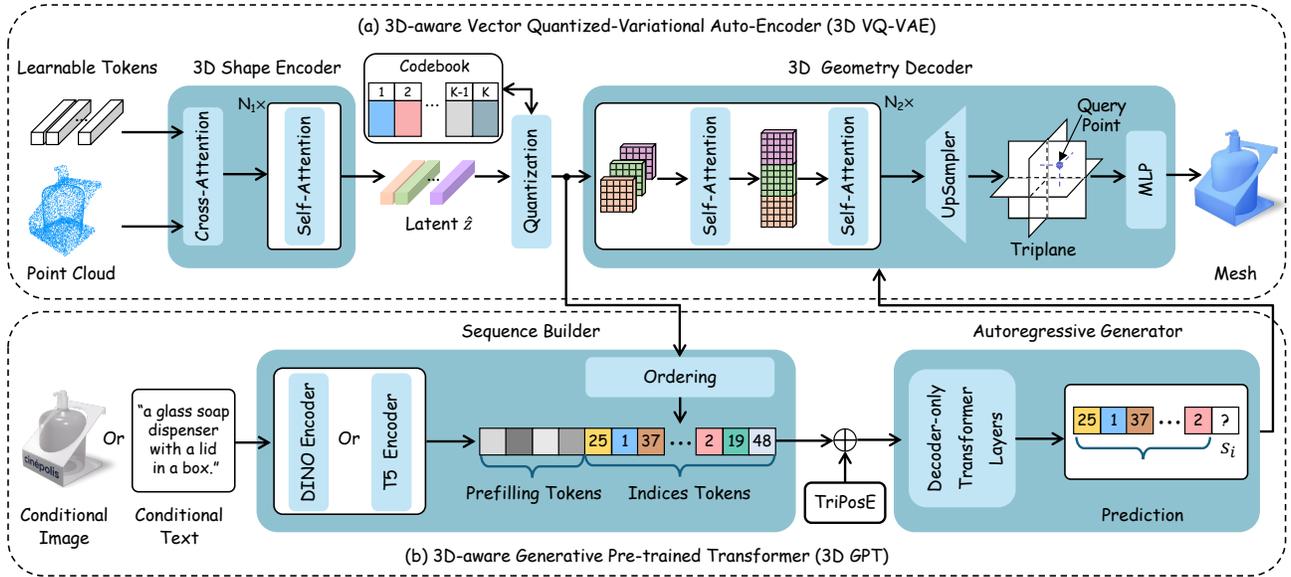
Figure 2. Overall architecture of the proposed TAR3D framework. (a) 3D VQ-VAE first encodes the point cloud uniformly sampled from 3D meshes into a set of learnable tokens in the triplane latent space. Then, these continuous triplane features are quantized as discrete embeddings from a trainable codebook. Next, these quantized representations are deformed twice, along with two self-attention modules in several attention layers to achieve feature enhancement in each plane and information interaction among the three planes. Subsequently, the triplane features are upsampled to a higher resolution for fine-grained geometry details. Finally, the query point features sampled from this triplane are fed to an MLP network for their occupancy predictions. (b) 3D GPT first organizes the triplane indices from the pre-trained codebook of 3D VQ-VAE into a sequence, in which the indices within each plane are placed in a raster scan order and the indices at the same positions of the three planes in an adjacent order. Then, the prompt features are employed as the prefilling token embedding of the sequence for conditional 3D object generation. Next, this sequence is modeled by multiple decoder-only transformer layers via next-part prediction. By querying the codebook, the predicted index sequence can be transformed into triplane features to synthesize 3D objects.

generation [52], music generation [19, 20], and speech gesture generation [2]. VQ-GAN [22] proposes to promote VQ-VAE by incorporating adversarial loss during the training process. In this paper, we build a 3D-aware VQ-VAE to quantize the triplane features of 3D shapes and acquire discrete geometric parts for autoregressive 3D generation.

2.3. GPT

Generative Pre-trained Transformer (GPT) [50] originates in the field of natural language processing (NLP). Based on the decoder-only transformer architecture, it autoregressively generates text sequences according to the next-token-prediction paradigm. This series of works [1, 3, 13, 36, 83] with groundbreaking reasoning ability and incredible scalability continue to emerge, revolutionizing language generation. Inspired by these achievements, many researchers have attempted to transfer this scheme to image generation. For example, Parti [76] proposes a pathways autoregressive text-to-image model, which regards image generation as sequence-to-sequence modeling for high-fidelity images. LlamaGen [59] verifies that vanilla autoregressive models, *e.g.,* Llama [63], can achieve state-of-the-art image generation performance without inductive biases on visual signals if scaling properly. Emu3 [69] achieves excellent per-

formance in both generation and perception tasks by tokenizing images, text, and videos into a discrete space. Besides, AutoSDF [47] learns a 'non-sequential' autoregressive shape prior for 3D completion, reconstruction, and generation. More recently, several works like MeshGPT [56], MeshAnything [11], and MeshXL [8] also explore to autoregressively generate the faces of 3D meshes with GPT.

In contrast to these methods, our TAR3D quantizes the triplane representations of 3D shapes into discrete geometric parts and uses the GPT model to generate 3D objects in a next-part prediction manner.

3. Methodology

Fig. 2 illustrates the overall architecture of our TAR3D framework. Our goal is to migrate the promising learning and multimodal unification capabilities of GPT [1] to conditional 3D object generation. However. existing methods [8, 11, 56] quantizing the mesh faces suffer from excessively long sequences, limiting their applications for high-quality 3D assets. In this work, we propose to represent the 3D shape information of meshes with triplane latent representations whose feature maps are associated with three-axis planes, *i.e.,* XY, YZ, and XZ. These triplane features can be quantized by a trainable codebook, resulting in a

fixed length sequence, regardless of the number of faces.

3.1. 3D VQ-VAE

To quantize the triplane representation of 3D shapes into discrete embeddings and synthesize 3D objects, we develop a 3D VQ-VAE, including a 3D shape encoder, a quantizer, and a 3D geometry decoder.

3D shape encoder is designed to acquire compact and robust latent representations of 3D objects for fine-grained geometry information. Given a 3D object, we first uniformly sample high-resolution point clouds from its surface. To enhance the expressive power of the point clouds, the corresponding normals are also included in the point cloud representation that is denoted as $P \in \mathbb{R}^{B \times N_p \times (3+3)}$, where B is the batch size, and N_p is the number of points. Then, we apply the Fourier positional [61] encoding to the point cloud representation to capture high-frequency details.

Inspired by previous works for point cloud understanding [29, 78], we employ a transformer-based architecture consisting of a cross-attention layer and N_1 self-attention layers to extract the latent features of the 3D point cloud. More precisely, the point cloud information is injected into a series of learnable query tokens that is denoted as $e \in \mathbb{R}^{B \times (3 \times h \times w) \times d_e}$, where h and w are the height and width of the triplane feature maps respectively, and d represents the channel number of these learnable tokens, via the cross-attention layers. After that, the representational ability of these tokens is enhanced by the following self-attention layers, resulting in triplane latent representations, *i.e.*, $\hat{z} \in \mathbb{R}^{B \times (3 \times h \times w) \times d_z}$.

Quantizer is introduced to represent the continuous triplane features with the embeddings z_q from a learnable, discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}^{d_q}$. Specifically, we first use a linear layer to project the continuous features to the same channel number with the codebook embeddings, yielding features $\tilde{z} \in \mathbb{R}^{B \times (3 \times h \times w) \times d_q}$. Then, an element-wise quantization between each spatial code of these features and its closest codebook entry is performed as follows:

$$z_q := \left(\arg\min_{z_k \in \mathcal{Z}} ||\tilde{z}_{ij} - z_k|| \right) \in \mathbb{R}^{B \times (3 \times h \times w) \times d_q}. \quad (1)$$

3D geometry decoder aims to reconstruct the 3D neural field in a high quality with the quantized discrete features, *i.e.*, z_q, as input. Inspired by recent works in video generation [24, 28] and avatar generation [68], we achieve plane information interaction by feeding z_q to N_2 attention layers consisting of two feature deformations and self-attention operations. In particular, the plane axis is ignored by being reshaped into the batch axis, allowing the first self-attention to process each plane independently. The plane axis is recovered and the features of the three planes are concatenated along the height dimension, enabling the second self-attention to model the information interaction across dif-
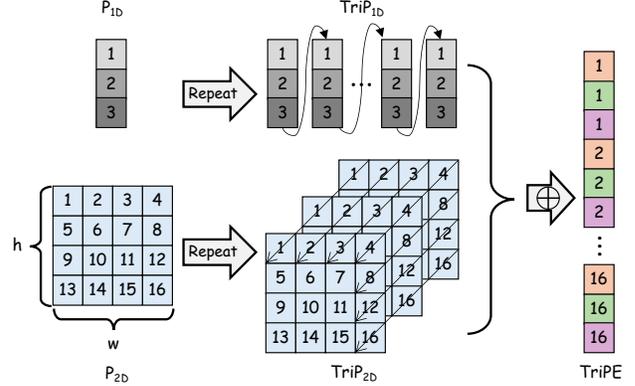

Figure 3. Diagrammatic details of our TriPE designed for the positional encoding of triplane sequence. We represent the positional information with numbers and simplify the number of tokens in 2D encoding for ease of presentation.

ferent planes. We also upsample the triplane features to a higher resolution following the operation in Direct3D [72]. Given a set of query points consisting of voxel points and near-surface points in the 3D field, their features can be sampled from the yielding triplane via bilinear interpolation. The query point features are transformed to their occupancy values via a Multi-Layer Perceptron (MLP).

3.2. 3D GPT

To model the constituents of a 3D object in an autoregressive manner, we propose a 3D GPT, which consists of three components, *i.e.*, a sequence builder, a TriPE position encoding, and an autoregressive generator.

Sequence builder. Based on the pre-trained 3D VQ-VAE, the 3D shapes are encoded as discrete triplane features, which can be represented with their indices in the codebook. Subsequently, these indices are transformed into a sequence according to some ordering rules. Considering the triplane representation is composed of three correlated feature maps, we organize the indices within each plane in a raster scan order and the indices at the same positions of the three planes in the adjacent orders. To achieve conditional 3D generation, the prompts are encoded as the prefilling token embedding of the sequence.

TriPE is a 3D position encoding strategy tailored for the triplane index sequence. As shown in Fig. 3, it is a fusion of 2D position encoding and 1D position encoding based on the Rotary Position Embedding (RoPE) [57]. We denote the RoPE for a 2D feature map with height h and width w as $P_{2D} \in \mathbb{R}^{h \cdot w}$, and the RoPE for a 1D sequence with 3 tokens as $P_{1D} \in \mathbb{R}^3$. Note that the channel dimension is removed for ease of description. To preserve the 2D spatial information in the axis-aligned feature planes for the triplane index sequence, we repeat the unit element of P_{2D} three times and place the two newly emerged elements adjacent to their original element. We denote this 2D position encoding for

triplane indices as $\text{TriP}_{2D} \in \mathbb{R}^{3 \cdot h \cdot w}$. Meanwhile, we repeat the three elements in P_{1D} for $h \times w$ times to highlight the difference of the three feature maps, yielding a 1D position encoding for triplane indices denoted as $\text{TriP}_{1D} \in \mathbb{R}^{3 \cdot h \cdot w}$. Finally, we calculate the TriPE by performing element-wise addition of TriP_{2D} and TriP_{1D}.

Autoregressive generator. After tokenizing the triplane indices into a sequence $s \in \{0, ..., K-1, K\}^{3 \cdot h \cdot w}$, along with the prompts c and custom positional encoding TriPE, 3D object generation can be formulated as an autoregressive next-index prediction. To be specific, the decoder-transformer layers learn to predict the distribution of possible next indices, which can be written as follows:

$$p_\theta(s|c) = \prod_t p_\theta(s_t|s_{<t}, c), \tag{2}$$

where t is the time step in the generation process, c is a conditional image or text embedding, and p_θ denotes the decoder-transformer layers with parameters θ.

3.3. Optimization Details

Corresponding to the overall architecture in Fig. 2, the optimization process of our TAR3D framework can also be divided into two stages, *i.e.,* 3D VQ-VAE optimization and 3D GPT optimization.

To train our 3D VQ-VAE in an end-to-end manner, we employ the Binary Cross-Entropy (BCE) loss as the optimization objective for reconstructing 3D objects. This process can be formalized as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{x \in \mathbb{R}^3} \left[\text{BCE}\left(\hat{\mathcal{O}}(x), \mathcal{O}(x)\right) \right], \tag{3}$$

where $\hat{\mathcal{O}}(\cdot)$, and $\mathcal{O}(\cdot)$ are the predicted occupancy value and ground-truth occupancy value of the query point. For the codebook learning of quantizer, the training loss can be formulated to minimize the difference between the original features and the quantified features:

$$\mathcal{L}_{cb} = ||sg(\tilde{z}) - z_q||_2^2 + \beta ||\tilde{z} - sg[z_q]||, \tag{4}$$

where $sg[\cdot]$ denotes the stop-gradient operation [4], and β is a weight hyperparameter to balance the two-part losses, which is set as $\beta = 0.25$ by default. Finally, our 3D VQ-VAE is optimized by minimizing:

$$\mathcal{L}_{3dvqvae} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cb}\mathcal{L}_{cb}, \tag{5}$$

where λ_{rec} and λ_{cb} are the weights of reconstruction optimization and codebook optimization, respectively.

The optimization objective of our 3D GPT is to maximize the log-likelihood of the triplane index sequence. As a result, its training loss can be expressed as follows:

$$\mathcal{L}_{3dgpt} = - \sum_{t=1}^{3 \cdot h \cdot w} \log \left(p_\theta(s_t|s_{<t}, c) \right). \tag{6}$$

4. Experiments

4.1. Experiment Setup

Datasets. To examine the effectiveness of our TAR3D, we conduct experiments on two benchmark 3D datasets, *i.e.,* ShapeNet [6], Objaverse [17], and an out-of-domain dataset, *i.e.,* Google Scanned Object(GSO) [21]. Specifically, the ShapeNet dataset provides 52,472 manufactured meshes covering 55 categories. We follow the splits from 3DILG [77], where 48,597 samples are used for training, 1,283 for validation, and 2,592 for testing. Inspired by previous works on data filtering [34, 80], we score the rendered normal maps of 800,000 meshes in the Objaverse dataset and obtain about 100,000 geometry objects. Moreover, 1,000 samples are randomly selected for performance evaluation, and the remaining ones are employed for model training. In addition, the GSO dataset contains about 1000 real-world 3D scans, which are utilized to further validate the generalization of our method. For each 3D asset in these two datasets, we adopt the rendered images and textual descriptions from ULIP [75] to build the prompt system. We uniformly select 4 images from all the rendered images, whose top 1 captions are utilized as the text prompts.

Metrics. We evaluate the performance of the methods from two aspects: 2D visual quality and 3D geometric quality. For the 2D visual evaluation, we compare the novel views rendered from the synthesized 3D mesh with the ground truth views based on a set of common metrics, including Peak Signal-to-Noise Ratio (PSNR), Perceptual Loss (LPIPS) [81], Structural Similarity (SSIM) [70], and CLIP [51] score. For the 3D geometric evaluation, we compare the point clusters that are randomly sampled from the generated meshes and the ground-truth meshes. Following the protocol in previous works [39, 73], we employ the chamfer distance value and the F-Score with a threshold of 0.02 as the primary evaluation metrics.

Implementation details. In our 3D VQ-VAE, the number of point clouds input to the 3D shape encoder, *i.e.,* N_p, is 81,920. The number of the self-attention layer in the 3D shape encoder, *i.e.,* N_1 is 8. The height and width of the triplane feature maps, *i.e.,* h and w, are both set to 32. The channel numbers of the learnable tokens, and the triplane features, *i.e.,* d_e and d_z, are set to 768 and 16. The size and channel number of the codebook, *i.e.,* K and d_q, are set to 16,384 and 8, respectively. Moreover, the number of the attention layers in the 3D geometry decoder is set as N_2=6. The triplane features are upsampled to a resolution of 256×256, in which 20,480 uniformly sampled voxel points and 20,480 near-surface points are employed as query points for occupancy supervision. The hyperparameters in Eqn. (5) are set as λ_{rec}=1 and λ_{cb}=0.1. We employ the CosineAnnealing scheduler [43], where the learning rate is initialized to $1e-4$ and gradually decays over time. We
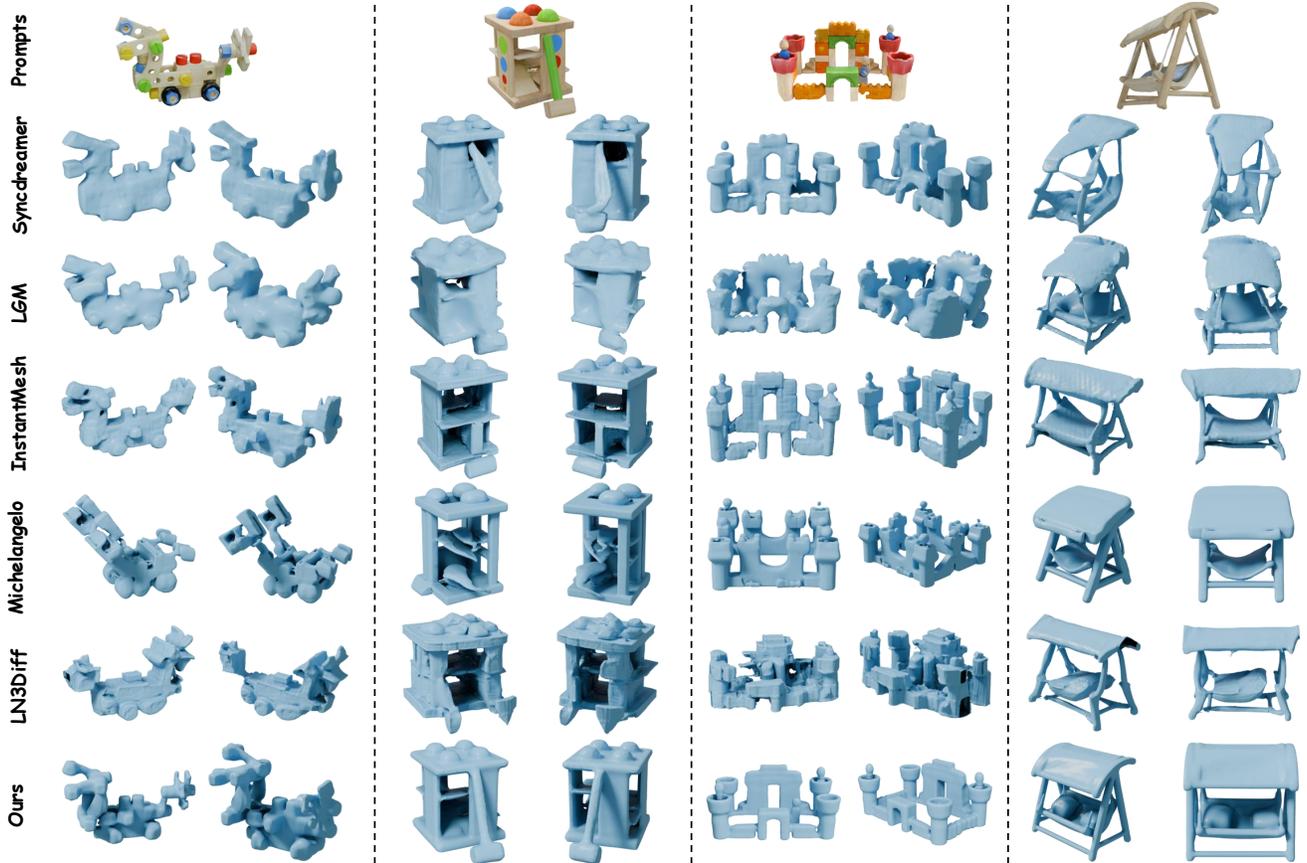
Figure 4. Visual comparisons of the 3D meshes generated by our TAR3D and recent multiview-based models, *i.e.,* Syncdreamer, LGM and InstantMesh, and 3D native approaches,*i.e.,* Michelangelo and LN3DIff, for image-to-3D object generation. Given the same input images from the GSO dataset, our TAR3D can produce 3D assets with superior geometric details against other baselines.

adopt the AdamW optimizer [44] to train the 3D VQ-VAE with a total batch size of 128 for 100K steps on 8 NVIDIA A100 GPUs.

In our 3D GPT, we adopt the pre-trained DINO [79](ViT-B16) and FLAN-T5 XL [16] to encode the conditional images and text prompts respectively. As far as the decoder-only transformer, we follow the GPT-L setting of LLama-Gen [59], which consists of 24 transformer layers with a head number of 16 and a dimension of 1024. We employ the AdamW optimizer with a learning rate of $1e-4$ and a total batch size of 80 to train our 3D GPT for 100K steps. In addition, the classifier-free guidance (CFG) [26] scale of 7.5 is also introduced in the autoregressive inference to improve geometry quality and image/text-3D alignment.

4.2. Qualitative Evaluation

Image-to-3D. We first provide visualization comparisons between our TAR3D with recent state-of-the-art models for image-to-3D generation, including three multiview-based methods like SyncDreamer [41], InstantMesh [73], OpenLRM [25], and LGM [62] and three 3D native

approaches like Shap-E [31], Michelangelo [86] and LN3Diff [33]. As shown in the garden-swing sample of Fig. 4, the methods based on mulit-view synthesis may produce 3D objects with discontinuous geometry parts or even incorrect objects, which are attributed to inconsistencies in their views used for reconstruction. Meanwhile, the LEGO sample shows that 3D diffusion methods are prone to generating noisy 3D objects, possibly due to anomalies in the conditional denoising process. In contrast to these approaches, our TAR3D model, equipped with the powerful learning capabilities of GPT scheme, can synthesize high-quality 3D objects with superior geometric details.

Text-to-3D. We compare our TAR3D with other cutting-edge text-to-3D approaches, including Diffusion-sdf [35], SDFusion [15], Shap-E [31], Fantasia3D [7], and Michelangelo [86]. As shown in Fig. 5, these baseline methods are susceptible to failure in various cases, either generating poor quality 3D objects, like the result of Michelangelo for "An advanced fighter Jet", or not matching the given textual descriptions, like the result of Fantasia3D for "a bed in the loft of the bunk bed". Different from them, our TAR3D can
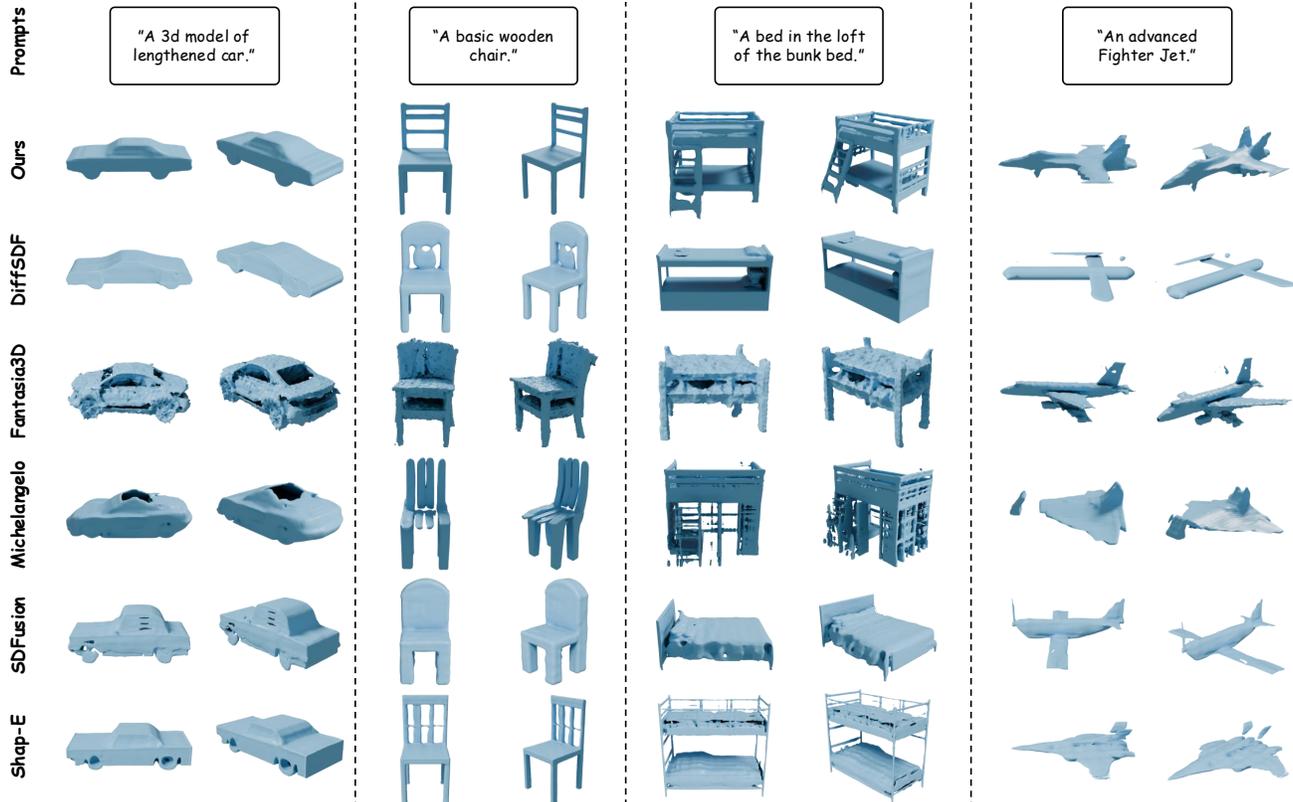
Figure 5. Qualitative comparisons of our TAR3D with recent cutting-edge methods for the text-to-3D object generation task. The 3D mesh assets created by our TAR3D are semantically more faithful to the given textual prompts compared to previous approaches.

Table 1. Quantitative comparisons of our TAR3D model with recent state-of-the-art image-to-3D generation methods on 2D visual quality and 3D geometric quality. '↑': the higher the value, the better the performance, '↓': the lower the better.

Methods	Shap-E [31]	SyncDreamer [41]	Michelangelo [86]	InstantMesh [73]	LGM [62]	TAR3D (Ours)
PSNR ↑	10.991	11.269	11.928	11.560	11.363	13.626
SSIM ↑	0.702	0.706	0.734	0.721	0.714	0.763
Clip-Score ↑	0.834	0.837	0.864	0.847	0.841	0.868
LPIPS ↓	0.325	0.320	0.278	0.303	0.317	0.216
Chamfer Distance ↓	0.156	0.158	0.117	0.137	0.149	0.066
F-Score ↑	0.163	0.178	0.226	0.179	0.172	0.303

create significantly more plausible 3D assets through GPT autoregression driven by text prefilling embeddings.

4.3. Quantitative Evaluation

In this subsection, we use the mixed evaluation set of ShapeNet and Objaverse to quantitatively measure the performance of our TAR3D model against recent 3D object generation methods in terms of 2D visual quality and 3D geometric quality. Considering the diversity of 3D objects generated from text descriptions, we conduct experiments on image-to-3D tasks to ensure an accurate comparison, as shown in Tab. 1,. In particular, all candidate approaches employ the same images as their conditional input to generate 3D meshes. For the 2D evaluation, we render 20 views with

224×224 resolution for each mesh and compare the yielding normal maps with the ones of corresponding ground-truth views. For the 3D evaluation, we compare the generated meshes with the ground truth meshes by uniformly sampling 16K point clouds from their surfaces in an aligned cube coordinate system of $[-1, 1]^3$. Our TAR3D model surpasses other cutting-edge 3D object generation methods across all 2D and 3D metrics by a large margin. These experimental results further demonstrate the superiority of our TAR3D over existing methods.

4.4. Ablation Studies

Evaluation on 3D VQ-VAE. We evaluate the 3D reconstruction capability of our 3D VQ-VAE, which serves as a

Table 2. Ablation studies on the reconstruction capability of the 3D VQ-VAE.

	3D VAE	3D VQ-VAE
Chamfer Distance ↓	0.018	0.016
F-Score ↑	0.811	0.822

Table 3. Ablation studies on the PII design of the 3D-VQVAE.

	w/o PII	w/ PII
Chamfer Distance ↓	0.023	0.016
F-Score ↑	0.661	0.822

Table 4. Ablation studies on triplane size (*i.e.,* sequence length) of the 3D GPT.

Triplane Size	$3\times16\times16$	$3\times32\times32$	$3\times48\times48$
Chamfer Distance ↓	0.157	0.066	0.062
Inference Time ↓	17.7 s	67.6 s	143.9 s

Figure 6. Ablation experiments on the effectiveness of our TriPE.

Figure 7. Visual comparisons of our TAR3D variants with different triplane sizes or sequence lengths.

foundation for generating high-quality 3D assets. For reference, we provide the performance of the VAE counterpart from which our 3D VQ-VAE is derived. To obtain a well-performing tokenizer for autoregressive generation, we adjust the training strategies until the performance of our 3D VQ-VAE is on par with or better than that of the VAE counterpart. The experiment results shown in Tab. 2 demonstrate the reconstruction effectiveness of our 3D VQ-VAE.

Plane information interaction in 3D VQ-VAE. We analyze the importance of plane information interaction (PII) achieved by the feature deformation and attention mechanism in our 3D VQ-VAE. As shown in Tab. 3, the variant without PII design achieves 0.661 f-score in 3D reconstruction. With the incorporation of our PII designs, this score significantly increased to 0.822. These experiments indicate that our PII designs contribute to the decoding process from the latent features to 3D objects.

TriPE positional encoding in 3D GPT. We investigate two positional encoding strategies for the sequence modeling in 3DGPT. The first strategy is the 1D Rotary Position Embedding (RoPE) [57] matching the sequence length, and the other is the proposed TriPE. As can be observed in Fig. 6, the 1D RoPE is prone to lose important geometric details of the objects in the input images. In contrast, our well-designed TriPE can preserve as much 3D spatial information as possible, thereby generating 3D objects that are geometrically more faithful to the prompts.

Triplane Size. We ablate the triplane size with $3\times16\times16$, $3\times32\times32$, and $3\times48\times48$ to explore the impact of sequence length on the performance and efficiency. Note that the $3\times32\times32$ widely adopted in 3D reconstruction is our default setting. As shown in Tab. 4, it is the best setting for efficiency and performance trade-offs, which is also demon-
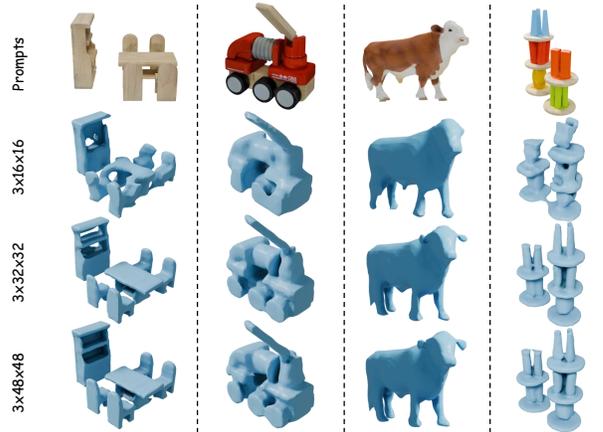
strated by the visual comparisons in Fig. 7

5. Conclusions & Future Work

In this paper, we propose a novel framework named TAR3D to generate high-quality 3D assets via the "next-token prediction" paradigm derived from multimodal large language models. To achieve this, we first develop a 3D VQ-VAE, in which the 3D shapes are encoded into the triplane latent space and quantized as discrete embeddings from a trainable codebook. Equipped with the well-designed feature deformation and attention mechanism, these quantized features are utilized to reconstruct fine-grained geometries in a neural occupancy field. Then, we adopt the codebook indices of these discrete representations to form the sequence for autoregressive modeling. Driven by the prefilling prompt embeddings, our 3D GPT, along with the 3D spatial information from the proposed TriPE, can generate high-quality 3D objects part by part.

We believe it is promising to improve 3D object generation under the autoregressive setting. There are a few foreseeable directions to explore: 1) Scaling law. We will collect more 3D data with prompts from other datasets, *e.g.,* Objaverse-XL [18], 3D-FUTURE [23], and scale the GPT model. 2) Sequence formulation. We will explore the sequence formulation manner of triplane indices for more efficient generative modeling.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3

[2] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022. 3

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 3

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 5

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 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. 2, 5

[7] 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*, pages 22246–22256, 2023. 1, 2, 6

[8] Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Wei Cheng, Yijun Fu, Fukun Yin, Yanru Wang, Zhibin Wang, Chi Zhang, et al. Meshxl: Neural coordinate field for generative 3d foundation models. *arXiv preprint arXiv:2405.20853*, 2024. 1, 3

[9] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in neural information processing systems*, 32, 2019. 2

[10] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. *arXiv preprint arXiv:2312.04424*, 2023. 2

[11] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024. 3

[12] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *arXiv preprint arXiv:2411.06558*, 2024. 2

[13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 3

[14] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26576–26586, 2025. 2

[15] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 6

[16] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 6

[17] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[18] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[19] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 3

[20] Sander Dieleman, Aaron Van Den Oord, and Karen Simonyan. The challenge of realistic music generation: modelling raw audio at scale. *Advances in neural information processing systems*, 31, 2018. 3

[21] 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. 2, 5

[22] 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. 3

[23] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 8

[24] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 4

[25] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 2, 6

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[27] 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. 2

[28] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 4

[29] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 4

[30] Jiayi Ji, Haowei Wang, Changli Wu, Yiwei Ma, Xiaoshuai Sun, and Rongrong Ji. Jm3d & jm3d-llm: Elevating 3d representation with joint multi-modal cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[31] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 6, 7

[32] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[33] 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 *European Conference on Computer Vision*, pages 112–130. Springer, 2025. 1, 2, 6

[34] 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. 5

[35] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12642–12651, 2023. 6

[36] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 3

[37] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023. 2

[38] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[39] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, et al. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 5

[40] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2

[41] 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. 2, 6, 7

[42] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[43] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[45] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024. 1

[46] 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

[47] 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. 3

[48] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[49] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[50] Alec Radford. Improving language understanding by generative pre-training, 2018. 3

[51] 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*. PMLR, 2021. 5

[52] 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

[53] 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

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

[55] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1, 2

[56] 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. 1, 3

[57] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4, 8

[58] Jia-Mu Sun, Tong Wu, and Lin Gao. Recent advances in implicit representation-based 3d shape generation. *Visual Intelligence*, 2(1):9, 2024. 2

[59] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2, 3, 6

[60] Minkui Tan, Qi Chen, Zixiong Huang, Qi Wu, Yuanqing Li, and Jiaqiu Zhou. Auto-3d-house design from structured user requirements. *Machine Intelligence Research*, 22(2):368–385, 2025. 2

[61] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 2020. 4

[62] 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. 2, 6, 7

[63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

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

[65] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[66] Haowei Wang, Jiayi Ji, Tianyu Guo, Yilong Yang, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Nice: improving panoptic narrative detection and segmentation with cascading collaborative learning. *arXiv preprint arXiv:2310.10975*, 2023. 2

[67] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[68] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 4

[69] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[70] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[71] 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. 2

[72] 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 preprint arXiv:2405.14832*, 2024. 1, 2, 4

[73] 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. 1, 2, 5, 6, 7

[74] Qun-Ce Xu, Tai-Jiang Mu, and Yong-Liang Yang. A survey of deep learning-based 3d shape generation. *Computational Visual Media*, 9(3):407–442, 2023. 2

[75] 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. 5

[76] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 3

[77] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *Advances*

in Neural Information Processing Systems, 35:21871–21885, 2022. 5

[78] 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, 4

[79] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 6

[80] 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, 5

[81] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[82] Tianyu Zhang, Guocheng Qian, Jin Xie, and Yang Jian. Fastpci: Motion-structure guided fast point cloud frame interpolation. In *ECCV*, 2024. 2

[83] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15465–15474, 2021. 3

[84] 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. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 19531–19540, 2024. 2

[85] Xuying Zhang, Yupeng Zhou, Kai Wang, Yikai Wang, Zhen Li, Shaohui Jiao, Daquan Zhou, Qibin Hou, and Ming-Ming Cheng. Ar-1-to-3: Single image to consistent 3d object generation via next-view prediction. *arXiv preprint arXiv:2503.12929*, 2025. 2

[86] 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. 2, 6, 7

[87] Zuo-Liang Zhu, Beibei Wang, and Jian Yang. GS-ROR: 3D Gaussian splatting for reflective object relighting via sdf priors. *arXiv preprint arXiv:2406.18544*, 2024. 2

[88] Zuo-Liang Zhu, Jian Yang, and Beibei Wang. Gaussian splatting with discretized sdf for relightable assets. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2025. 2