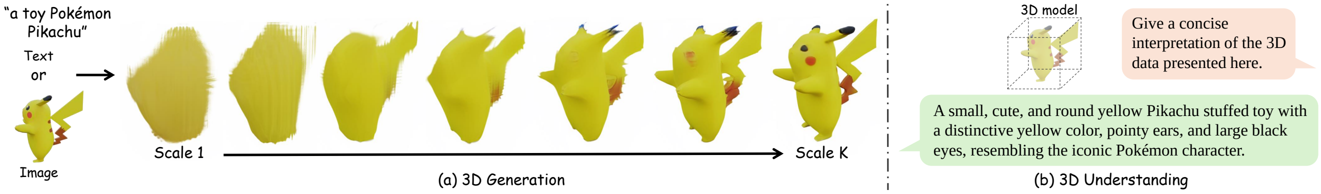
SAR3D: Autoregressive 3D Object Generation and Understanding via Multi-scale 3D VQVAE

Yongwei Chen[1] Yushi Lan[1] Shangchen Zhou[1] Tengfei Wang[2] Xingang Pan[1]
[1]S-Lab, Nanyang Technological University
[2]Shanghai Artificial Intelligence Laboratory
https://cyw-3d.github.io/projects/SAR3D/



Give a concise interpretation of the 3D data presented here.

A small, cute, and round yellow Pikachu stuffed toy with a distinctive yellow color, pointy ears, and large black eyes, resembling the iconic Pokémon character.

Figure 1. Our method, SAR3D, proposes a comprehensive framework for 3D generation and understanding via autoregressive modeling. For (a) 3D generation, given a single image or text prompt, SAR3D generates multi-scale 3D objects in an autoregressive manner. For (b) 3D understanding, SAR3D-LLM can interpret a 3D model and provide a detailed description.

Abstract

*Autoregressive models have demonstrated remarkable success across various fields, from large language models (LLMs) to large multimodal models (LMMs) and 2D content generation, moving closer to artificial general intelligence (AGI). Despite these advances, applying autoregressive approaches to 3D object generation and understanding remains largely unexplored. This paper introduces Scale AutoRegressive 3D (*SAR3D*), a novel framework that leverages a multi-scale 3D vector-quantized variational autoencoder (VQVAE) to tokenize 3D objects for efficient autoregressive generation and detailed understanding. By predicting the next scale in a multi-scale latent representation instead of the next single token,* SAR3D *reduces generation time significantly, achieving fast 3D object generation in just* 0.82 *seconds on an A6000 GPU. Additionally, given the tokens enriched with hierarchical 3D-aware information, we finetune a pretrained LLM on them, enabling multimodal comprehension of 3D content. Our experiments show that* SAR3D *surpasses current 3D generation methods in both speed and quality and allows LLMs to interpret and caption 3D models comprehensively.*

1. Introduction

Autoregressive models have achieved remarkable success in various domains, including large language models (LLMs) [1, 5, 14, 50, 72, 73], 2D generation [66, 71, 88], and large multimodal models (LMMs) [2, 17, 67], marking significant strides toward artificial general intelligence (AGI). By predicting the next token [1] or scale [71], autoregressive models are trained using a simple cross-entropy loss and share similar architectures. This commonality allows them to easily benefit from the optimizations the community has developed over the years for LLMs. Nevertheless, there has been limited exploration of how this next-token/scale prediction approach can be applied to 3D object generation and understanding.

Previously, the scarcity of 3D data pushed researchers to rely on pretrained 2D diffusion models [58] as a prior to generate 3D objects via multi-view score distillation sampling (SDS) loss [52]. Following this, alternative approaches [27, 68] have focused on training feed-forward 3D reconstruction models for fast 3D reconstruction, enabled by large-scale 3D object datasets like Objaverse [15, 16]. These methods are capable of generating 3D assets in mere seconds. More recently, native 3D generative models [29, 34, 48, 93] have emerged, attempting to sample 3D assets from noise under various conditions (*e.g.*, text or image). However, as most of these models rely on diffusion-based methods, they suffer from slow inference times. In parallel, mesh-based generative models [11, 63] attempt to generate 3D topology using autoregressive predictions, but they are limited in detail and require slow, face-by-face predictions. For 3D understanding, some studies [19, 25, 85] have attempted to finetune LLMs on 3D data to interpret the 3D world. However, these methods primarily use 3D point cloud representations, which are limited in capturing fine object details.

In light of the immense potential of autoregressive next-token prediction paradigm and their underexplored status in 3D generation and understanding, we pose an important question: *Can autoregressive models be effectively applied to achieve both fast 3D object generation and detailed understanding?* Addressing this challenge requires a 3D tokenizer capable of encapsulating detailed information about 3D objects into compact tokens, as well as an efficient schedule for autoregressive prediction.

In this work, we propose *Scale AutoRegressive 3D* (SAR3D), a framework that leverages autoregressive models for both fast object generation and comprehensive understanding. Central to SAR3D is a multi-scale 3D vector-quantized variational autoencoder (VQVAE) capable of tokenizing 3D objects into hierarchical levels of tokens. These multi-scale tokens facilitate next-scale prediction training, significantly reducing the steps required for 3D generation compared to diffusion models and traditional next-token prediction methods. Furthermore, the tokens, enriched with 3D-aware information, are naturally compatible with LLM fine-tuning for detailed 3D understanding.

Specifically, our SAR3D introduces a multi-scale 3D VQVAE to encode multiview RGB images, along with their corresponding depth and camera parameters, into a multi-scale latent triplane representation. For 3D generation, we train an autoregressive model to predict the next scale of this latent triplane based on previous scales, conditioned on a single image or text prompt. By predicting the next scale instead of the next single token, our approach significantly reduces generation time, achieving 3D object generation in only 0.82 seconds on an A6000 GPU. For 3D understanding, we use truncated scale tokens from our 3D multi-scale VQVAE to finetune a pretrained LLM, enabling it to process multimodal inputs that combine text and 3D tokens. Notably, our finetuned LLM can interpret 3D tokens encoded by our VQVAE as well as those generated by our autoregressive model, supporting both 3D captioning and simultaneous generation and understanding.

Experiments show that SAR3D surpasses existing 3D generation methods in both speed and quality, and our VQVAE enables LLMs to generate detailed captions for 3D objects . Our key technical contributions are as follows:

- We introduce SAR3D, a framework designed for both fast 3D object generation and detailed 3D understanding.
- For 3D generation, our method utilizes a next-scale prediction approach for both text-to-3D and single-image-to-3D, achieving faster generation with higher quality compared to existing methods.
- For 3D understanding, we leverage truncated scale tokens generated by our 3D multi-scale VQVAE to finetune a pretrained LLM, enabling it to interpret and describe 3D models, and showcasing the potential of our approach in multimodal applications.

2. Related Works

3D Generative Models. With the success of 2D diffusion models [23, 65], their adaptation for 3D generation has been widely explored. Score distillation sampling [10, 12, 52, 69, 77] leverages these 2D models to distill 3D content, yet it encounters challenges such as costly optimization, mode collapse, and the Janus problem. More recent approaches adopt a two-stage pipeline, generating multi-view images first [42, 61, 62, 78] and then reconstructing 3D structures through feed-forward processes [26, 70, 84]. Although promising, these methods are constrained by the quality of multi-view image generation, which often lacks view consistency [41] and fails to scale to higher resolutions [61]. Additionally, this two-stage setup limits 3D editing capabilities due to the absence of a 3D-aware latent space.

To overcome these limitations, native 3D diffusion models [34, 35, 37, 75, 91–93] have been introduced, offering high-quality, efficient, and scalable 3D generation. Native 3D diffusion pipelines use a two-stage training process: first encoding 3D objects into a VAE latent space [32, 33], followed by applying a latent diffusion model on the resulting codes. However, diffusion-based 3D generation is slow during inference, and its latent space cannot be easily renovated for 3D understanding. In parallel, mesh generative models [11, 63] generates 3D topology through autoregressive prediction. However, they lack details and require slow per-face prediction. In this study, we show that our autoregressive SAR3D achieves efficient sampling, better quality, and can naturally be used for 3D understanding by cascading a Large Language Model.

Autoregressive Visual Generation. Pioneered by PixelCNN [59], researchers have proposed to generate images as pixel sequences. Early research VQVAE [74] and VQGAN [18] further quantize image patches into discrete tokens and employ a transformer to learn the autoregressive priors, similar to language modeling [1]. The following research further improves its sampling speed [8] and tokenization efficiency [89]. To further improve the reconstruction quality, RQVAE [36] proposed multi-scale quantization and VAR [71] transforms it into next-scale prediction and significantly enhances the sampling speed. Parallel efforts are spent on scaling up autoregressive models on text-conditioned visual generation task [38, 56, 66, 76]. In 3D area, though some preliminary works [47, 90] studied 3D autoregressive modeling on toy dataset [7] without textures, the research on autoregressive 3D generation over large scale 3D dataset [15, 16] is missing.

Large Multimodal Models. Inspired by the great success of large language models (LLMs) [6, 72, 73], large multimodal models (LMMs) are proposed to comprehend and generate a wide range of information beyond text-based data. Two prominent paradigms exist to train the models in an end-to-end strategy: training the model from scratch [45] or aligning pre-trained LLMs and unimodal encoders [2, 39]. The second strategy typically involves a two-stage process: alignment of the unimodal encoder with LLM's feature space, and instruction-based fine-tuning. Following up works also extend the LMMs to 3D understanding, specifically on point cloud [25, 53, 85, 86]. However, point clouds significantly ignore the details of the given 3D inputs. Here, we demonstrate that our 3D VQVAE can connect with an LLM for a detailed 3D understanding.

3. Preliminaries

3.1. Multi-scale Visual Autoregressive Generation

VAR [71] presents a multi-scale visual modeling method for image generation, shifting from "next-token prediction" to "next-scale prediction", which significantly improves the inference speed of autoregressive models. Given an encoded feature map $f \in \mathbb{R}^{h \times w \times C}$ of an input image I, VAR quantizes f into K multi-scale token maps $R = (r_1, r_2, ..., r_K)$ at an increasingly higher resolution $h_k \times w_k$, with r_K matches the resolution of the input feature map f. The autoregressive likelihood reads as:

$$p(r_1, r_2, ..., r_K) = \prod_{k=1}^{K} p(r_k | r_1, r_2, ..., r_{k-1}), \quad (1)$$

where each autoregressive unit $r_k \in [V]^{h_k \times w_k}$ is the token map at scale k, and the sequence $(r_1, r_2, ..., r_{k-1})$ serves as the prefix for r_k. To tokenize the input image I to multi-scale discrete token maps R for the learning of next-scale prediction, VAR proposes a multi-scale VQVAE with a multi-scale quantizer $\mathcal{Q}(\cdot)$:

$$f = \mathcal{E}(I), \quad R = \mathcal{Q}(f), \quad (2)$$

where I denotes the raw image and $\mathcal{E}$ is the image encoder. This quantization process will map f to a sequence of multi-scale token maps by looking up the nearest code [74] in codebook $Z \in \mathbb{R}^{V \times C}$:

$$z_k^{(i,j)} = \left(\arg\min_{v \in [V]} \left\| \text{lookup}(Z, v) - r_k^{(i,j)} \right\|_2 \right) \in [V], \quad (3)$$

where lookup(Z, v) means taking the v-th vector in codebook Z. To train the quantized autoencoder, Z is looked up by every $z_k(i,j)$ to get $\hat{f}$, the approximation of original f. Then a new image $\hat{I}$ is reconstructed using the decoder $D(\cdot)$ given $\hat{f}$:

$$\hat{f} = \text{lookup}(Z, z), \quad \hat{I} = \mathcal{D}(\hat{f}). \quad (4)$$

Once fully trained, the autoencoder $\{\mathcal{E}, \mathcal{Q}, \mathcal{D}\}$ will tokenize the incoming images for training the unidirectional autoregressive model.

3.2. PointLLM for Point Cloud Understanding

Given multi-modal sentences containing both point clouds $P \in \mathbb{R}^{n \times d}$ and text, where n is the number of points and d is the dimension of each point, PointLLM [85] aims to perform 3D point cloud understanding by finetuning a pretrained large language model [13, 73]. It consists of three main components: a pretrained point cloud encoder Γ_{pe} (*e.g.*, Point-BERT [90]), a projector Γ_{proj} and a pretrained large language model backbone Γ_{llm}. The Γ_{pe} and Γ_{proj} project P to a point cloud token sequence $Z_p \in \mathbb{R}^{m \times c'}$, where m is the total number of tokens and c' is the projected dimension of the point tokens. The final mixed sequence of tokens $Z_m = (z_1, z_2, ..., z_l) \in \mathbb{R}^{l \times c}$ consist of both point tokens Z_p and text tokens Z_t:

$$Z_p = \Gamma_{\text{proj}}(\Gamma_{\text{pe}}(P)), \quad Z_m = \text{Concat}(Z_p, Z_t), \quad (5)$$

where Z_t is obtained by tokenizer of Γ_{llm}, and Concat$(\cdot)$ means concatenation of two vectors. The LLM backbone Γ_{llm} is a GPT-style Transformers [5], which accepts a sequence of previous multi-modal tokens $Z_{<i} = (z_1, \dots, z_{i-1})$ and predict the next token:

$$z_i = \Gamma_{\text{llm}}(Z_{<i}). \quad (6)$$

The finetuning process has two stages. In the first stage, it freezes $\{\Gamma_{\text{pe}}, \Gamma_{\text{llm}}\}$ and finetunes Γ_{proj} to align point features with the text token space. In the second stage, it freezes Γ_{pe} and finetunes $\{\Gamma_{\text{llm}}, \Gamma_{\text{proj}}\}$ together.

4. Method

In this section, we present SAR3D for high-quality 3D object generation and detailed understanding. First, we introduce a multi-scale 3D vector-quantized variational autoencoder (VQVAE) in Sec. 4.1, which tokenizes the input 3D model into multi-scale tokens. Fig. 2 illustrates the design of our 3D VQVAE. Next, in Sec. 4.2, with the whole sequence of different scales of the feature tokens, we train an autoregressive model to perform the next scale prediction given a single image or text prompt, which is only supervised by simple cross-entropy loss. Finally, in Sec. 4.3, we explore using *truncated scales* of the whole sequence to fine-tune a pretrained LLM [13, 72, 73] to handle multimodal input sequence containing both 3D and text tokens, thereby enabling the understanding of the input 3D model. Fig. 3 illustrates our 3D generation and understanding pipeline. Different from other methods [80] that train different encoders for generation and understanding, we train a single VQVAE and use the whole and truncated sequence for generation and understanding, respectively. Details of our SAR3D are shown below.

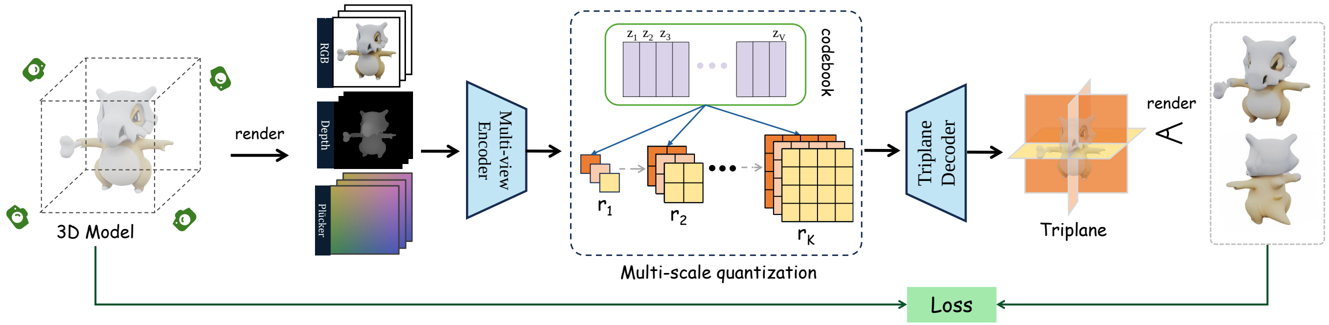


Figure 2. **Overview of Multi-scale VQVAE**. Given a 3D model, we leverage multi-view RGB-D(epth) renderings and Plücker embeddings as the input to our multi-view encoder $\mathcal{E}$. The encoder predicts a continuous feature map that is then quantized by the multi-scale quantizer $\mathcal{Q}$, giving $R = (r_1, r_2, \ldots, r_K)$ of latent tri-plane features. Each code of different scales share the same codebook. The triplane decoder then converts the quantized latent triplane features into the triplane representation through a plane-wise manner. The predicted triplane is multi-view supervised with the ground truth image, depth, and normal.

4.1. Multi-scale 3D VQVAE

As demonstrated by previous studies [4, 8, 34, 57], the key to high-quality visual generation lies in a compact latent space achieved by a specially designed variational autoencoder [32, 74]. To achieve *both fast 3D generation and detailed understanding*, we propose a multi-scale 3D VQVAE that maps a given 3D object into a discrete multi-scale latent space. To encode a 3D model, we leverage its multi-view posed RGB-D renderings as input. This approach offers a comprehensive representation of the 3D structure and enables compatibility with existing architectures [81].

Specifically, the input of VQVAE is a set of multi-view renderings of the 3D object from 6 views. Each rendering $M = (I, \Delta, \pi)$ captures essential 3D attributes, representing the object from a specific viewpoint: the RGB image $I \in \mathbb{R}^{H\times W\times 3}$, the depth map $\Delta \in \mathbb{R}^{H\times W}$, and the corresponding camera pose π. To standardize these 3D attributes, we transform the camera pose π into Plücker coordinates [64], expressed as $\mathbf{p}_i = (\mathbf{o} \times \mathbf{d}_{u,v}, \mathbf{d}_{u,v}) \in \mathbb{R}^6$, where $\mathbf{o}_i \in \mathbb{R}^3$ denotes the camera origin, $\mathbf{d}_{u,v} \in \mathbb{R}^3$ is the normalized ray direction, and $\times$ represents the cross product. Consequently, the Plücker embedding of the camera π is represented as $\mathbf{P} \in \mathbb{R}^{H\times W\times 6}$. The final representation is formed by channel-wise concatenating these elements, resulting in $\tilde{M} = [I \oplus \Delta \oplus \mathbf{P}] \in \mathbb{R}^{H\times W\times 10}$, where $\oplus$ denotes concatenation.

To maintain both geometry and texture details of M, similar to LN3Diff [34], we encode the inputs through a multi-view convolutional encoder [62, 70]. For better 3D awareness, the latent space is designed to be a latent triplane [34, 82] $f \in \mathbb{R}^{3\times h\times w\times C}$. Besides, this representation also has spatial inductive bias and is compatible with the scale and interpolation design in VAR [71]. After encoding, f is interpolated to different scales and quantized using latent triplane quantization layer $\mathcal{Q}$:

$$f = \mathcal{E}(\tilde{M}), \quad R = \mathcal{Q}(f), \tag{7}$$

where $\mathcal{E}$ is the encoder of our VQVAE and $R = (r_1, r_2, ..., r_K)$ is the scale sequence and $r_k \in \mathbb{R}^{3\times h_k\times w_k\times C}$, where each sub latent plane $r_k^i \in \mathbb{R}^{h_k\times w_k\times C}$ is independently quantized and interpolated over the shared codebook Z. Please refer to the *Supp Mat* for more quantization and interpolation details.

Afterward, the decoder $\mathcal{D}$ decodes discrete scales R to triplane and render multiple views to calculate the reconstruction loss. To balance training stability and mesh extraction quality, we first train our model on volume rendering [46] with the loss reads as:

$$\mathcal{L} = \lambda_{\text{render}}\mathcal{L}_{\text{render}} + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}, \tag{8}$$

where $\mathcal{L}_{\text{render}}$ combines mean absolute error (MAE) and perceptual loss [94] between the rendered RGB-D images with masks and ground truth, $\mathcal{L}_{\text{VQ}}$ includes both encoding error and commitment loss [74], and $\mathcal{L}_{\text{GAN}}$ serves as the adversarial loss to encourage a perceptually rich latent space. λ_* are the corresponding loss weights.

To facilitate 3D mesh extraction, we further finetune the model to the hybrid representation Flexicubes [60, 84] with the extra $\mathcal{L}_{\text{flex}}$ loss:

$$\mathcal{L}_{\text{flex}} = \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{9}$$

where $\mathcal{L}_{\text{normal}}$ is MAE loss between rendered normal and ground truth, $\mathcal{L}_{\text{reg}}$ is regularization term for Flexicubes parameters [60]. λ_* are the corresponding loss weights. Similar to LATTE3D [83], we only fine-tune the decoder of the VQVAE in this stage for stabilized training.

4.2. 3D Generation via Multi-scale Autoregressive Modeling

SAR3D Transformer. We illustrate our generation framework in Fig. 3. Similar to VAR [71], we use standard GPT-style transformer [5] with AdaLN layer [51], with specific layer design following the simple rule of scaling law [30].

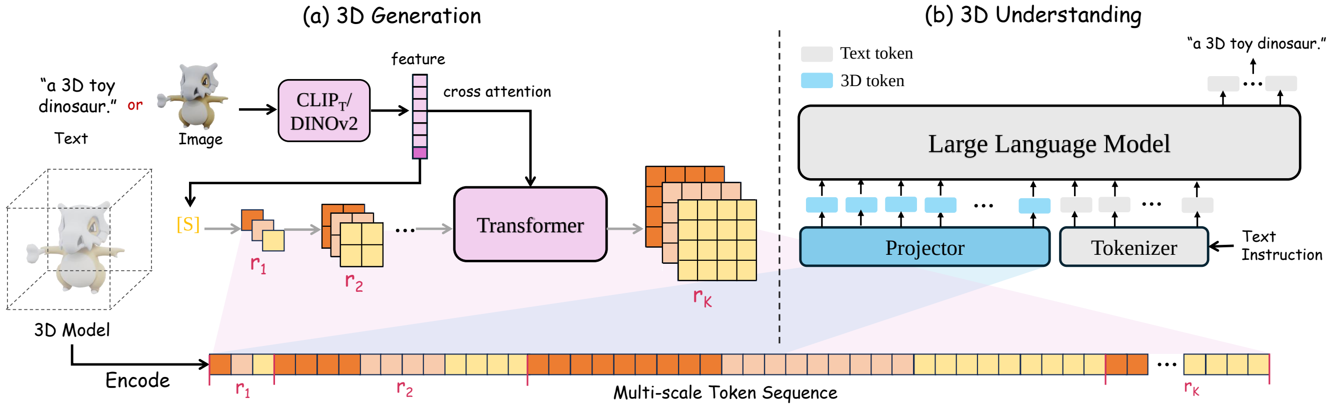


Figure 3. **Overview of 3D Generation and 3D Understanding.** Given a 3D model, our 3D VQVAE encodes it into multi-scale discrete tokens for both 3D generation and understanding. In (a) 3D Generation, text or a single image is encoded by CLIP_T or DINOv2, and the encoded condition features are integrated into the decoder-only transformer via cross attention. The transformer then causally predicts each scale of the latent triplane. In (b) 3D Understanding, truncated 3D tokens are first processed with an MLP projector. The large language model receives a multimodal sequence of text and 3D tokens and generates a detailed caption describing the input 3D model.

We adopt tri-plane latent for autoregressive prediction with Eq. 1, where different latent plane r_k^i is differentiated with the corresponding learnable positional embeddings.

Conditional 3D Generation. Unlike feed-forward 3D reconstruction models [27, 70] that map input image into 3D, we achieve flexible multimodal 3D generation by introducing diverse conditions, as shown in Fig. 3. For text conditions, we use CLIP_T [55] ViT-L text encoder and inject the text embeddings into the autoregressive models through cross attention. For the image-conditioned model, we use DINOv2 [49] ViT-L to extract local patch features and send them into the autoregressive model through pre-cross-attention block [27], which empirically yields better performance. Besides local patch features, we also leverage the pooled out feature of CLIP_T/DINOv2 as the start token of the sequence. Please refer to the *Supp Mat* for more details of our transformer blocks.

Classifier-free Guidance. As first proposed in diffusion models [23, 57], classifier-free guidance [22] (CFG) has shown effectiveness for improving generation quality and condition alignment. Therefore, we also enable CFG in our model by randomly dropping out 10% of the input condition by replacing it with a *null* unconditional embedding [51]. During inference, the logit r_g of each token is calculated by $r_g = r_u + s(r_c - r_u)$, given the conditional logit r_c and the unconditional logit r_u. s stands for the scale of the classifier-free guidance.

4.3. SAR3D-LLM for 3D Object Understanding

Since our 3D VQVAE model provides a comprehensive encoding of the given 3D object, it can be naturally extended to 3D object understanding. Following PointLLM [85], we align the latent space of our pre-trained 3D VQVAE to a large language model, *e.g*., LLaMA [72, 73]. As briefed in Sec. 3.2, the encoded 3D tokens R are projected to the language latent space, and concatenated with the text instruction tokens Z_t. Here, we directly use the output tokens from pre-trained SAR3D VQVAE to the projector Γ_proj. Since we only study the 3D captioning [43] task here, the instruction tokens Z_t is fixed to $\tilde{Z}_t$ which are tokenized from *"Give a concise interpretation of the 3D data presented here"*. The final framework, SAR3D-LLM, supports both detailed 3D captioning given 3D object and simultaneous 3D generation and captioning given text or image.

Moreover, a surprising observation here is that not all scales in R are required for 3D understanding training. Empirically, we use the *truncated scale latent codes* $\tilde{R} = (r_1, r_2, ..., r_{K-2})$ as the input to the LLM, which contains only *37.5%* of the overall tokens required for training 3D generation. The final multimodal tokens Z_m that serve as the input to the LLM reads as

$$Z_{proj} = \Gamma_\text{proj}(\tilde{R}), \quad Z_m = \text{Concat}(Z_{proj}, \tilde{Z}_t), \qquad (10)$$

where Z_{proj} is the projected 3D tokens, and $\text{Concat}(\cdot, \cdot)$ means concatenation. A similar observation is also mentioned in Janus [80], where different features are required for multimodal understanding and generation. Furthermore, unlike other 3D captioning approaches such as Cap3D [43, 44], which separately extracts captions from 8 multi-view renderings and require post-processing to merge them into a unified caption, our method efficiently generates a detailed caption with a single encoding step.

5. Experiments

Datasets. To train our model, we use renderings from G-Objaverse [16, 54] and select a high-quality subset of around $176K$ 3D instances, where each consists of 40 ran-

Table 1. **Quantitative Evaluation of Image-conditioned 3D Generation.** We evaluate the quality of both 2D rendering and 3D shapes. As shown below, the proposed method demonstrates strong performance across all metrics. Although LGM, a multi-view images-to-3D approach, achieves slightly better performance on FID, it falls short on more advanced image quality assessment metrics such as MUSIQ and has significantly worse 3D shape quality. For multi-view to 3D methods, we also include the number of input views (V=#). The latency time is all profiled on Tesla V100 architecture.

Method	FID↓	KID(%)↓	MUSIQ↑	COV(%)↑	MMD(‰)↓	Latency-V100 (s) ↓
Splatter-Image	48.80	3.65	30.33	37.66	30.69	0.83
OpenLRM	38.41	1.87	45.46	39.33	29.08	7.21
One-2-3-45 (V=12)	88.39	6.34	59.02	33.33	35.09	59.23
Lara (V=4)	43.74	1.95	39.37	39.33	28.84	11.93
CRM (V=6)	45.53	1.93	64.10	38.83	28.91	22.10
LGM (V=4)	19.93	0.55	54.78	50.83	22.06	3.87
Shap-E	138.53	11.95	31.51	61.33	19.17	9.54
LN3Diff	29.08	0.89	50.39	55.17	19.94	7.51
SAR3D-NeRF	22.55	0.42	67.24	71.50	15.24	1.64
SAR3D-Flexicubes	27.30	0.63	65.17	59.50	15.48	2.92

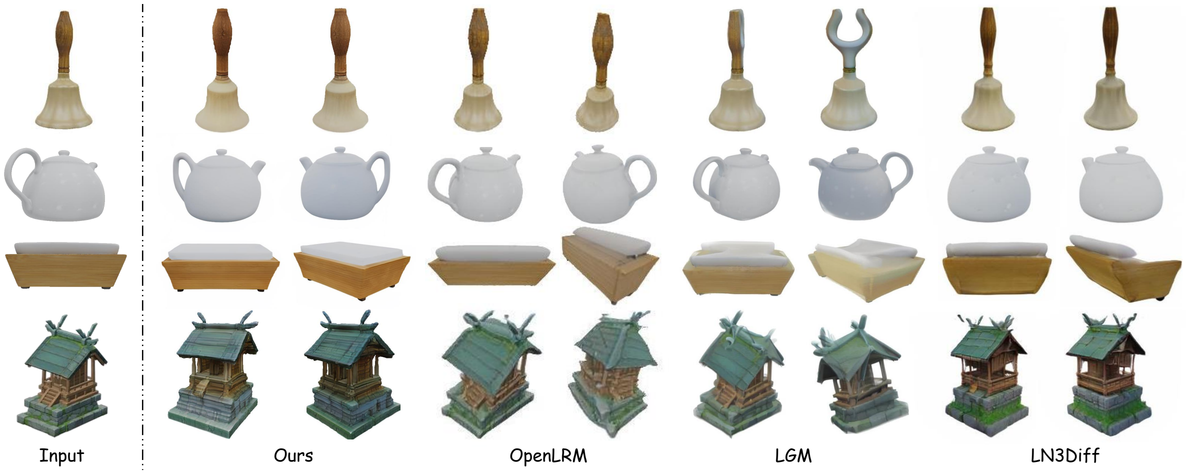


Figure 4. **Qualitative Comparison of Image-conditioned 3D Generation.** Here, we compare with the state-of-the-art 3D generative models under different categories. As visualized here, our method achieves superior 3D consistency across views and generates intact objects without distortion. For more comparisons with other methods, please refer to the *Supp Mat*.

dom views with RGB, normal, depth maps and camera poses. For text-conditioned generation and 3D understanding training, we use captions provided by 3DTopia [24]. For image-conditioned training, we select a random view of the corresponding 3D instance as the condition.

Implementation Details. In our multi-scale VQVAE, we use images with a resolution of $H = W = 256$ as input. The feature map is quantized across 10 scales, with sizes of $3 \times (1^2, 2^2, 3^2, 4^2, 5^2, 6^2, 8^2, 10^2, 13^2, 16^2)$. To enhance codebook utilization and stabilize 3D generation training, we follow [66, 87] by applying ℓ_2-normalization to codebook vectors, setting a low codebook vector dimension $C = 8$, and using a large codebook size $V = 16384$.

For 3D generation, we base our architecture on VAR [71], adding plane positional encoding for each plane. For text-conditioned generation, the model has 16 transformer blocks with 16 heads, while for image-conditioned generation, it has 24 transformer blocks with 16 heads. We use the AdamW optimizer with a learning rate of 10^{-4}. For 3D understanding, we utilize the Vicuna-7B [13] checkpoint of LLaMA [73], following PointLLM [85]. The training was conducted on 7 NVIDIA A100 GPUs for the multi-scale VQVAE with batch size 28, image-conditioned transformer with batch 63, text-conditioned transformer with batch size 52. For SAR3D-LLM, the stage-1 alignment is trained with batch size 140, and stage-2 with 112.

5.1. Single Image to 3D

We compare our SAR3D with three categories of methods: *single-image to 3D methods* (Splatter-Image [68], OpenLRM [20, 26]), *multi-view image to 3D methods* (One-2-3-45 [40], Lara [9], CRM [79], LGM [70]),

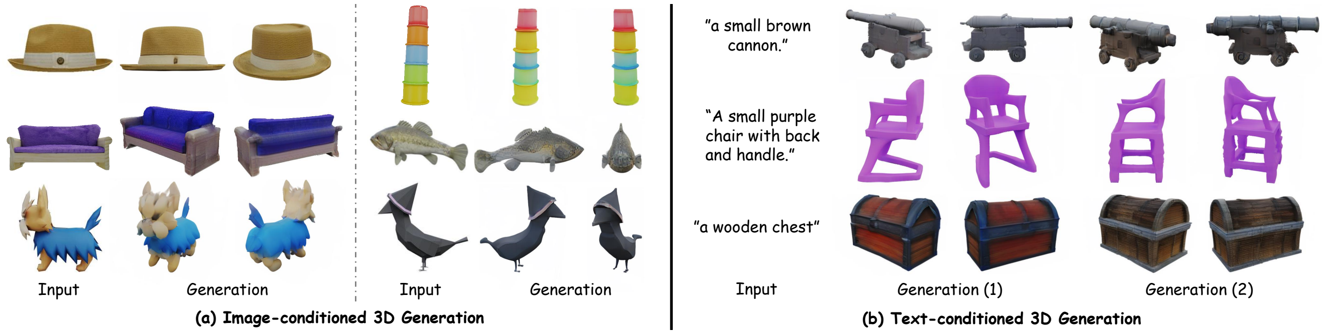

"a small brown cannon."
"A small purple chair with back and handle."
"a wooden chest"
Input
Generation
Input
Generation
(a) Image-conditioned 3D Generation
Input
Generation (1)
Generation (2)
(b) Text-conditioned 3D Generation

Figure 5. **More results of image and text conditioned 3D generation of SAR3D.**

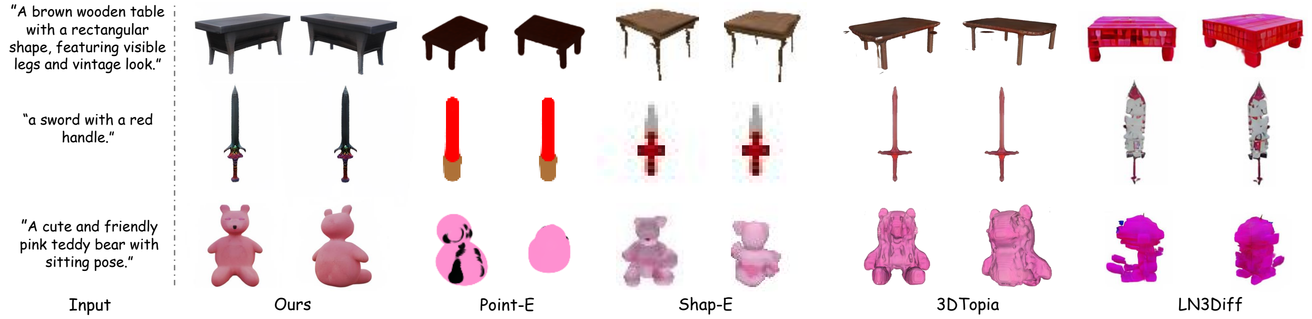

"A brown wooden table with a rectangular shape, featuring visible legs and vintage look."
"a sword with a red handle."
"A cute and friendly pink teddy bear with sitting pose."
Input
Ours
Point-E
Shap-E
3DTopia
LN3Diff

Figure 6. **Comparison of Text-conditioned 3D Generation**. We present text-conditioned 3D objects generated by SAR3D, displaying two views of each sample.Compared to baseline methods, our approach consistently yields better quality regarding geometry, texture, and text-3D alignment.

and *native 3D diffusion models* (Shap-E [29], LN3Diff-image [34]). Quantitatively, we benchmark rendering metrics with FID [21], KID [3], and MUSIQ [31, 95]. For 3D quality evaluation, we report Coverage Score (COV), and Minimum Matching Distance (MMD) score, as shown in Table 1. Our SAR3D demonstrates strong performance across all metrics.

Furthermore, we also profile the generation speed. This timing covers the complete process, from input image processing to mesh extraction. Thanks to efficient next-scale prediction, SAR3D achieves exceptionally fast generation speeds, achieving 0.82 seconds and 1.46 seconds respectively on a single A6000 GPU. Since other baseline methods are tested on Tesla V100 GPUs, we scale our results by a factor of 2 for fair comparison in Table 1.

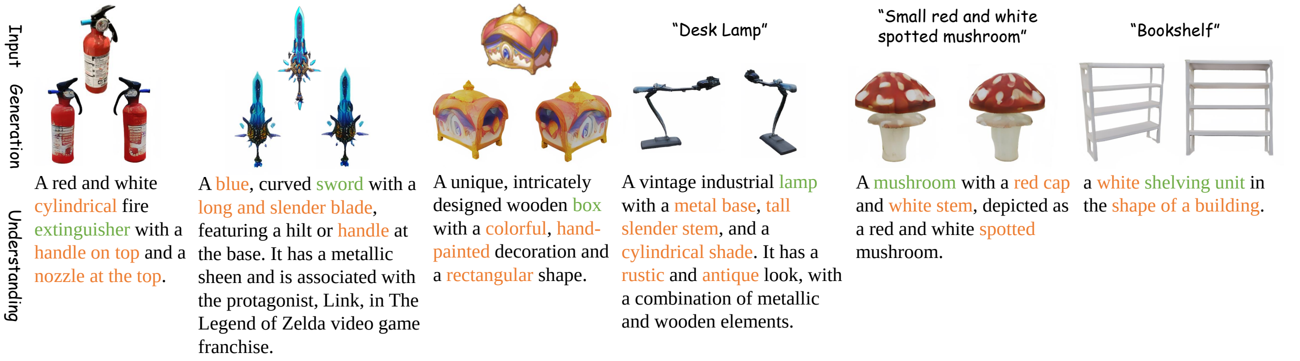
The qualitative comparisons between SAR3D and existing methods is also included in Fig. 4. Compared to *single-image to 3D methods* like OpenLRM[20], and *multi-view image to 3D methods* like LGM [70], our approach achieves better 3D consistency across views and reduces distortion in generated 3D objects. Compared to *native 3D diffusion models* like LN3Diff [34], SAR3D produces more complete 3D models. Additional qualitative results are shown in Fig. 5. For more comparisons with other methods, please refer to the *Supp Mat*.

5.2. Text to 3D

In addition to image-to-3D generation, SAR3D also supports the creation of high-quality 3D assets from text prompts. As shown in Fig. 5, SAR3D generates diverse and detailed 3D objects based on the same text input. For instance, in the first and second samples, SAR3D produces different shapes for the cannon barrel and chair base, while in the third sample, it varies the texture of a wooden chest. In Fig. 6, we compare our method with other text-to-3D generation approaches, including Point-E [48], Shap-E [29], 3DTopia [24], and LN3Diff [34]. Compared to these baselines, SAR3D achieves sharper visual results and better alignment with the input prompts. For example, in the second sample, SAR3D generates red patterns on the handle, closely matching the input text.. In contrast, Point-E [48] reverses the colors of the handle and blade, 3DTopia [24] produces a completely red sword, and Shap-E [29] yields a less detailed result.

5.3. 3D Captioning

3D Object Captioning. In this section, we present the results of our 3D understanding model applied to various 3D models. As shown in Fig. 8, given the prompt *"Give a concise interpretation of the 3D data presented here."*, SAR3D-LLM can generate both the correct category and



A red and white cylindrical fire extinguisher with a handle on top and a nozzle at the top.

A blue, curved sword with a long and slender blade, featuring a hilt or handle at the base. It has a metallic sheen and is associated with the protagonist, Link, in The Legend of Zelda video game franchise.

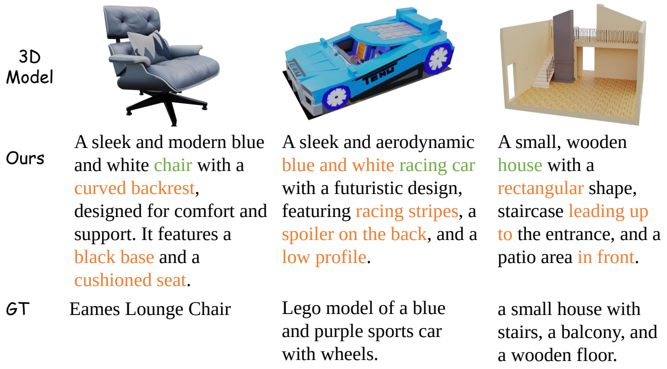
A unique, intricately designed wooden box with a colorful, hand-painted decoration and a rectangular shape.

A vintage industrial lamp with a metal base, tall slender stem, and a cylindrical shade. It has a rustic and antique look, with a combination of metallic and wooden elements.

A mushroom with a red cap and white stem, depicted as a red and white spotted mushroom.

a white shelving unit in the shape of a building.

Figure 7. **Simultaneous 3D Generation and Captioning**. Given a single image or text, SAR3D-LLM can generate both a 3D model and a descriptive caption for the model.



A sleek and modern blue and white chair with a curved backrest, designed for comfort and support. It features a black base and a cushioned seat.

A sleek and aerodynamic blue and white racing car with a futuristic design, featuring racing stripes, a spoiler on the back, and a low profile.

A small, wooden house with a rectangular shape, staircase leading up to the entrance, and a patio area in front.

Eames Lounge Chair

Lego model of a blue and purple sports car with wheels.

a small house with stairs, a balcony, and a wooden floor.

Figure 8. **3D Object Captioning**. Given a 3D model, SAR3D-LLM can generate captions that include both category and details.

fine details of the input 3D models. For example, in the chair case, SAR3D accurately describes the shape (curved backrest), colors (blue and white), and components (black base, cushioned seat), whereas the ground truth text lacks these details. Furthermore, our 3D tokens enable the LLM to capture the spatial relationships between different parts of the model. For instance, in the third column of Fig. 8, SAR3D uses phrases like "leading up to" and "in front of" to describe the spatial relationship between the staircase, entrance, and patio area, while the ground truth label merely lists these parts without capturing their spatial connections.

Simultaneous 3D Generation and Captioning. In addition to interpreting tokens encoded by our 3D VQVAE, SAR3D can also process 3D tokens generated by our autoregressive model to enable simultaneous 3D generation and captioning, as illustrated in Fig. 7. Given the condition input image or text, SAR3D not only generates the entire object but also detailed captions based on truncated scales of the generated 3D tokens. Notably, in text-conditioned generation and understanding, SAR3D generates additional details beyond those specified in the input text, resulting in accurate and comprehensive descriptions of the generated content.

6. Limitations

The first limitation is that, while SAR3D can generate high-quality 3D objects and detailed interpretations, it currently relies on two separate autoregressive models. Future work could focus on developing a truly multimodal model [80] capable of processing tokens that integrate both text and 3D information, producing both 3D and text outputs. Besides, the quality of the geometry and texture is limited by volume rendering. Using more efficient 3D represention [28] or cascaded generation [93] will further boost the overall quality. Finally, although our method demonstrates inherent scalability, its scaling behavior has not been thoroughly validated here due to limited resources. We believe that with more resources, our method has the potential to demonstrate favorable scaling laws in 3D generation and understanding.

7. Conclusion

In this work, we presented SAR3D, a novel framework that advances both fast 3D object generation and comprehensive 3D understanding through a multi-scale VQVAE and autoregressive modeling. By introducing a latent tri-plane next-scale prediction approach, we addressed the speed limitations of existing diffusion-based 3D generation methods, achieving sub-second generation times with high-quality results. Furthermore, our multi-scale VQVAE enables a pretrained LLM to process and interpret multimodal inputs by leveraging truncated scale 3D tokens, demonstrating the capability of LLMs for detailed 3D object captioning as well as simultaneous 3D generation and captioning. Experimental results underscore SAR3D's efficiency and effectiveness in 3D generation and understanding tasks, positioning it as a versatile tool for multimodal AI applications. Future research may further explore scalability and extend SAR3D's application to broader 3D content and multimodal understanding challenges.

Supplemental Materials for SAR3D: Autoregressive 3D Object Generation and Understanding via Multi-scale 3D VQVAE

Yongwei Chen[1] Yushi Lan[1] Shangchen Zhou[1] Tengfei Wang[2] Xingang Pan[1]

[1]S-Lab, Nanyang Technological University
[2]Shanghai Artificial Intelligence Laboratory

https://cyw-3d.github.io/projects/SAR3D/

A. Multi-scale quantization and interpolation

Similar to VAR [71], we employ quantization and interpolation in a residual design on the latent tri-plane feature map, as described in Algorithm 1 and Algorithm 2. In particular, they demonstrate that all scales share the same codebook, and each plane of the latent tri-plane is quantized independently based on the corresponding plane's previous scales. To upsample z_k^i to the resolution of $h_K \times w_K$, we utilize convolutional layers $\phi_k^i(\cdot)$. For interpolating z_k^i to resolution $h_K \times w_K$, we don't use any network.

Algorithm 1 Multi-scale 3D VQVAE Encoding

```
Require: multiview renderings $\tilde{M}$
Require: steps $K$, resolutions $(3, h_k, w_k)_{k=1}^{K}$
$f \leftarrow \mathcal{E}(\tilde{M})$, $R \leftarrow []$;
for $k = 1, \ldots, K$ do
    for $i = 1, \ldots, 3$ do
        $r_k^i \leftarrow \mathcal{Q}(\text{interpolate}(f, h_k, w_k))$
        $R \leftarrow \text{queue\_push}(R, r_k^i)$
        $z_k^i \leftarrow \text{lookup}(Z, r_k^i)$
        $z_k^i \leftarrow \text{interpolate}(z_k^i, h_K, w_K)$
        $f^i \leftarrow f^i - \phi_k^i(z_k^i)$
    end for
end for
return multi-scale latent tri-plane tokens $R$
```

B. Transformer blocks

The architecture of our transformer block for 3D generation is illustrated in Fig. S1. We utilize the CLIP text encoder or the DINOv2 image encoder to process text and image embeddings, respectively. The pooled tokens are then passed through an MLP to compute the scale and shift parameters for the multi-head self-attention and feedforward network (FFN) modules. Additionally, the feature vectors are incorporated into multi-head cross-attention blocks to facilitate cross-modal attention. To enhance the integration of cross-modal information into the model, similar to [35], we modify the structure of the transformer blocks by rearranging the order of self-attention and cross-attention in the text-conditioned and image-conditioned transformer blocks.

Algorithm 2 Multi-scale 3D VQVAE Reconstruction

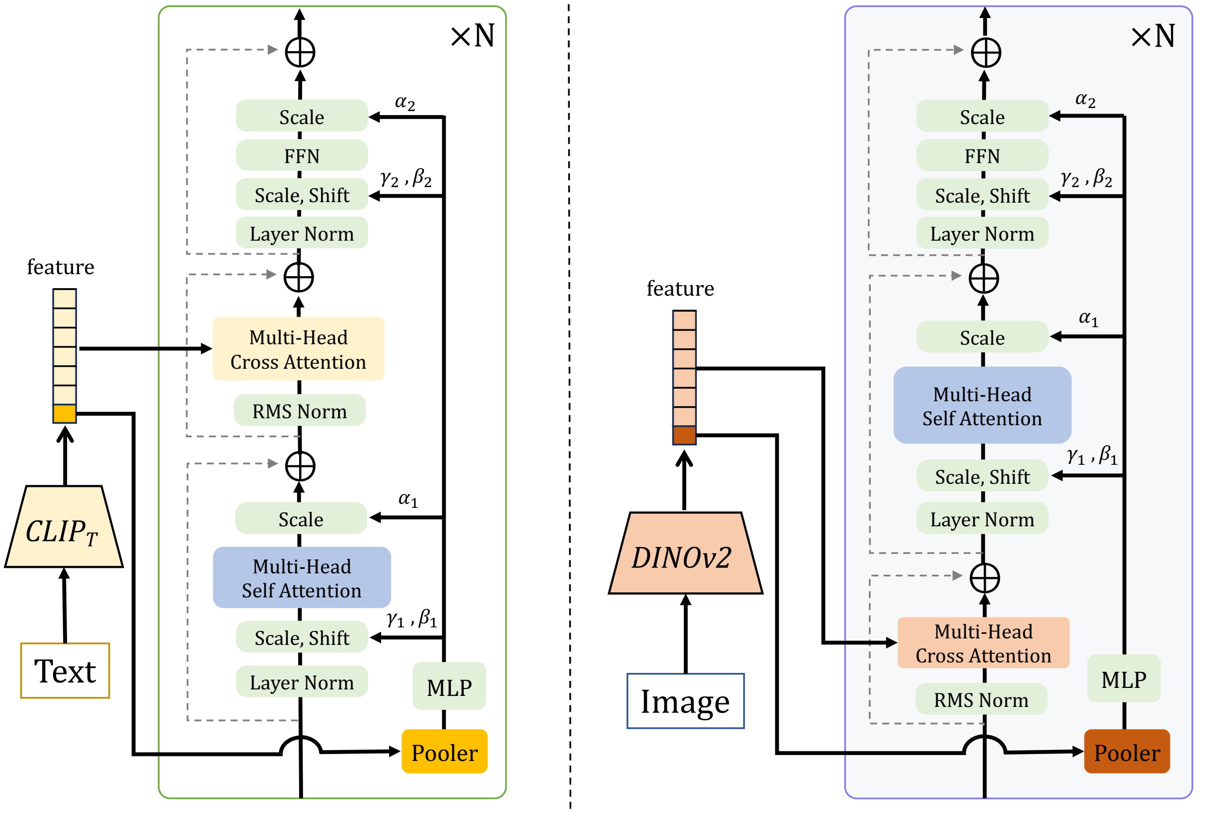
```
Require: multi-scale latent tri-plane token maps $R$
Require: steps $K$, resolutions $(3, h_k, w_k)_{k=1}^{K}$
$\hat{f} \leftarrow 0$
for $k = 1, \ldots, K$ do
    for $i = 1, \ldots, 3$ do
        $r_k^i \leftarrow \text{queue\_pop}(R)$
        $z_k^i \leftarrow \text{lookup}(Z, r_k^i)$
        $z_k^i \leftarrow \text{interpolate}(z_k^i, h_K, w_K)$
        $\hat{f}^i \leftarrow \hat{f}^i + \phi_k^i(z_k)$
    end for
end for
$\hat{T} \leftarrow \mathcal{D}(\hat{f})$
return reconstructed triplane representation $\hat{T}$
```

C. More 3D captioning results

Additional 3D captioning results are presented in Fig. S2. Given a 3D model, our SAR3D-LLM is capable of generating detailed captions. For instance, in the case of the *skateboard ramp*, our method can describe specific details about its shape, such as *curved, flat top, sloping bottom*, as well as its functionality, like *performing tricks and jumps.*

D. More image-to-3D comparison

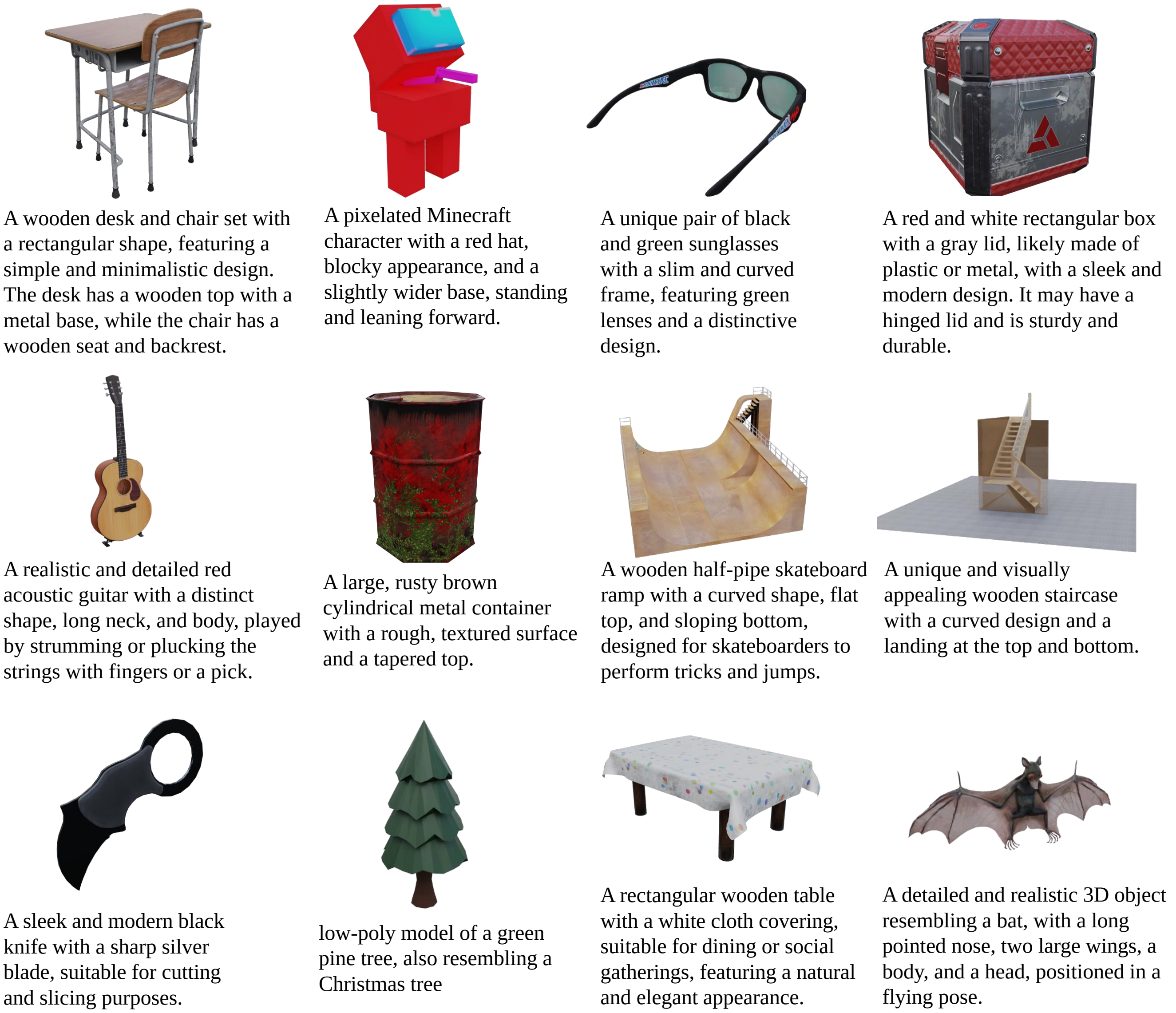
As illustrated in Fig. S3, we show more reults to compare our SAR3D with three categories of methods: *single-image to 3D methods* (Splatter-Image [68], OpenLRM [20, 26]), *multi-view image to 3D methods* (One-2-3-45 [40], Lara [9], CRM [79], LGM [70]), and *native 3D diffusion models* (Shap-E [29], LN3Diff-image [34]). Compared to baseline methods, our SAR3D generates intact, distortion-free results and delivers high-quality visual effects in both reference and novel views.



(a) Transformer Block (Text condition)

(b) Transformer Block (Image condition)

Figure S1. **Transformer Blocks in Our 3D Generation Transformer.** The CLIP text encoder (CLIP$_T$) or the DINOv2 image encoder processes text and image embeddings, respectively. The pooled tokens are passed through an MLP to compute the scale and shift parameters for the multi-head self-attention and feedforward network (FFN) modules. Additionally, feature vectors are incorporated into multi-head cross-attention blocks to enable cross-modal attention.



A wooden desk and chair set with a rectangular shape, featuring a simple and minimalistic design. The desk has a wooden top with a metal base, while the chair has a wooden seat and backrest.

A pixelated Minecraft character with a red hat, blocky appearance, and a slightly wider base, standing and leaning forward.

A unique pair of black and green sunglasses with a slim and curved frame, featuring green lenses and a distinctive design.

A red and white rectangular box with a gray lid, likely made of plastic or metal, with a sleek and modern design. It may have a hinged lid and is sturdy and durable.

A realistic and detailed red acoustic guitar with a distinct shape, long neck, and body, played by strumming or plucking the strings with fingers or a pick.

A large, rusty brown cylindrical metal container with a rough, textured surface and a tapered top.

A wooden half-pipe skateboard ramp with a curved shape, flat top, and sloping bottom, designed for skateboarders to perform tricks and jumps.

A unique and visually appealing wooden staircase with a curved design and a landing at the top and bottom.

A sleek and modern black knife with a sharp silver blade, suitable for cutting and slicing purposes.

low-poly model of a green pine tree, also resembling a Christmas tree

A rectangular wooden table with a white cloth covering, suitable for dining or social gatherings, featuring a natural and elegant appearance.

A detailed and realistic 3D object resembling a bat, with a long pointed nose, two large wings, a body, and a head, positioned in a flying pose.

Figure S2. **Additional 3D Captioning Results.** Our method generates detailed descriptions based on the input of 8 scales of latent tri-plane tokens.

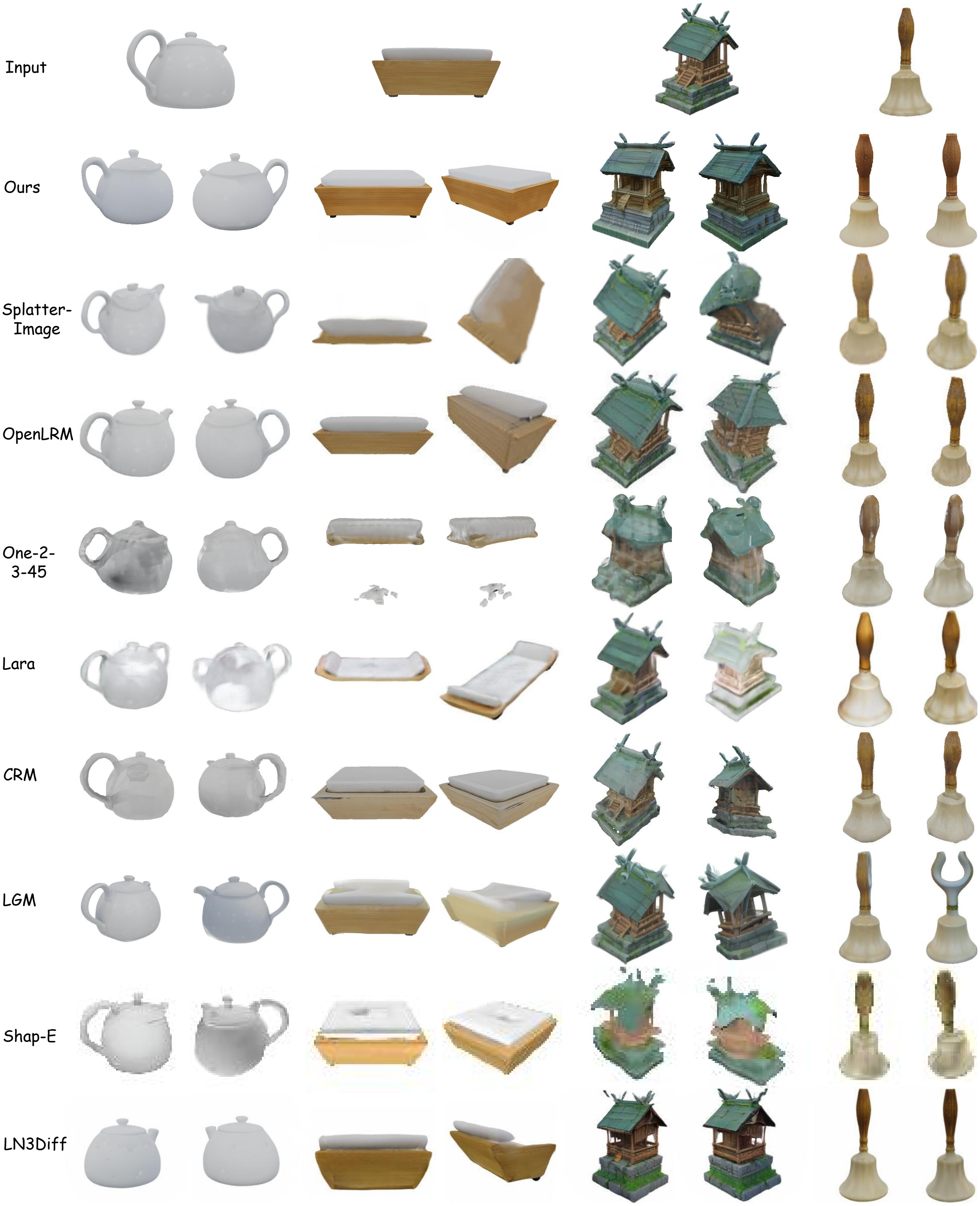


Figure S3. **More Comparisons of Image-to-3D Generation.** Our method consistently produces higher-quality 3D objects without distortion from a single image, excelling in both reference and novel views.

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, 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1, 3

[3] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 7

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 4

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 1, 3, 4

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. 2020. 2

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[8] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022. 2, 4

[9] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *ECCV*, 2024. 6, 9

[10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 2

[11] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, Guosheng Lin, and Chi Zhang. Meshanything: Artist-created mesh generation with autoregressive transformers, 2024. 1, 2

[12] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3D assets creation using spatially-aware diffusion guidance. *arXiv preprint arXiv:2403.12409*, 2024. 2

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3, 6

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 1

[15] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. *arXiv preprint arXiv:2307.05663*, 2023. 1, 2

[16] 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 *CVPR*, 2023. 1, 2, 5

[17] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *ICML*. PMLR, 2023. 1

[18] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[19] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 1

[20] Zexin He and Tengfei Wang. OpenLRM: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 6, 7, 9

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 7

[22] Jonathan Ho. Classifier-free diffusion guidance. In *NeurIPS*, 2021. 5

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2, 5

[24] Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Tengfei Wang, Liang Pan, Dahua Lin, and Ziwei Liu. 3DTopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, 2024. 6, 7

[25] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3D world into large language models. *NeurIPS*, 2023. 1, 3

[26] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao

Tan. Lrm: Large reconstruction model for single image to 3d. In *ICLR*, 2024. 2, 6, 9

[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. In *ICLR*, 2024. 1, 5

[28] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 8

[29] Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3D implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 7, 9

[30] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. 4

[31] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 7

[32] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013. 2, 4

[33] Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sovna Mokr'a, and Danilo Jimenez Rezende. NeRF-VAE: A geometry aware 3D scene generative model. *ICML*, 2021. 2

[34] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3D generation. In *ECCV*, 2024. 1, 2, 4, 7, 9

[35] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3d generation. 2024. 2, 9

[36] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022. 2

[37] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. CraftsMan: High-fidelity mesh generation with 3D native generation and interactive geometry refiner, 2024. 2

[38] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining, 2024. 2

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[40] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 6, 9

[41] 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, 2023. 2

[42] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *CVPR*, 2024. 2

[43] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3D captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 5

[44] Tiange Luo, Justin Johnson, and Honglak Lee. View selection for 3D captioning via diffusion ranking. *arXiv preprint arXiv:2404.07984*, 2024. 5

[45] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023. 3

[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. In *ECCV*, 2020. 4

[47] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. AutoSDF: Shape priors for 3D completion, reconstruction and generation. In *CVPR*, 2022. 2

[48] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A system for generating 3D point clouds from complex prompts, 2022. 1, 7

[49] 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. 5

[50] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022. 1

[51] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4, 5

[52] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *ICLR*, 2022. 1, 2

[53] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3D object understanding for embodied interaction. *arXiv preprint arXiv:2402.17766*, 2024. 3

[54] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. *arXiv preprint arXiv:2311.16918*, 2023. 5

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4, 5

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[59] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications. In *ICLR*, 2017. 2

[60] 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 Trans. Graph.*, 42(4), 2023. 4

[61] 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. In *arXiv*, 2023. 2

[62] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3D generation. *arXiv:2308.16512*, 2023. 2, 4

[63] 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 *CVPR*, 2023. 1, 2

[64] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *NeurIPS*, 2021. 4

[65] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[66] 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. 1, 2, 6

[67] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *ICLR*, 2023. 1

[68] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *arXiv*, 2023. 1, 6, 9

[69] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22819–22829, 2023. 2

[70] 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 *ECCV*, 2024. 2, 4, 5, 6, 7, 9

[71] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 2024. 1, 2, 3, 4, 6, 9

[72] 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. 1, 2, 3, 5

[73] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 2, 3, 5, 6

[74] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 3, 4

[75] 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. 2

[76] 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. 2

[77] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In *NeurIPS*, 2023. 2

[78] Zhenwei Wang, Tengfei Wang, Zexin He, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. Phidias: A generative model for creating 3d content from text, image, and 3d conditions with reference-augmented diffusion. *arXiv preprint arXiv:2409.11406*, 2024. 2

[79] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. CRM: Single image to 3D textured mesh with convolutional reconstruction model. In *ECCV*, 2024. 6, 9

[80] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 3, 5, 8

[81] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3D reconstruction. *arXiv preprint arXiv:2301.08247*, 2023. 4

[82] 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, 2024. 4

[83] Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. Latte3d: Large-scale amortized text-to-enhanced3d synthesis. *ECCV*, 2024. 4

[84] 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. 2, 4

[85] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024. 1, 3, 5, 6

[86] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning unified representation of language, image and point cloud for 3D understanding. *arXiv preprint arXiv:2212.05171*, 2022. 3

[87] 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. 6

[88] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *ICLR*, 2024. 1

[89] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *NeurIPS*, 2024. 2

[90] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3D point cloud transformers with masked point modeling. In *CVPR*, 2022. 2, 3

[91] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3D shape generation. In *NeurIPS*, 2022. 2

[92] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 42(4), 2023.

[93] 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 TOG*, 43(4):1–20, 2024. 1, 2, 8

[94] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[95] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 7