G3PT: Unleash the Power of Autoregressive Modeling in 3D Generation via Cross-Scale Querying Transformer

Jinzhi Zhang[1] , **Feng Xiong**[2] , **Guangyu Wang**[2,3] and **Mu Xu**[4]
[1]AMAP
[2]AMAP
[3]Tsinghua University
[4]AMAP
{wushou.zjz, xf250971, xumu.xm}@alibaba-inc.com, wanggy24@mails.tsinghua.edu.cn,

Abstract

Autoregressive transformers have revolutionized generative models in language processing and shown substantial promise in image and video generation. However, these models face significant challenges when extended to 3D generation tasks due to their reliance on next-token prediction to learn token sequences, which is incompatible with the unordered nature of 3D data. Instead of imposing an artificial order on 3D data, in this paper, we introduce G3PT – a scalable, coarse-to-fine 3D native generative model with *cross-scale vector quantization* and *cross-scale autoregressive modeling*. The key is to map point-based 3D data into discrete tokens with different levels of detail, naturally establishing a sequential relationship across a variety of scales suitable for autoregressive modeling. Remarkably, our method connects tokens globally across different levels of detail without manually specified ordering. Benefiting from this approach, G3PT features a versatile 3D generation pipeline that effortlessly supports the generation of 3D shapes under diverse conditional modalities. Extensive experiments demonstrate that G3PT achieves superior 3D generation quality and generalization ability compared to previous baselines. Most importantly, for the first time in 3D generation, scaling up G3PT reveals distinct power-law scaling behaviors.

1 Introduction

In recent years, the field of 3D shape generation has experienced significant advancements. One notable approach is the use of Large Reconstruction Models (LRMs) [Hong *et al.*, 2023; Tochilkin *et al.*, 2024], which convert images into 3D shapes through a pipeline that employs transformers [Vaswani *et al.*, 2017] to create and optimize implicit 3D representations with multi-view image supervision. Another line of approach extends 2D diffusion models [Rombach *et al.*, 2022] into the 3D domain, aiming to combine multi-view images into consistent 3D shapes using techniques such as sparse view reconstruction [Li *et al.*, 2023] and score distillation sampling [Poole *et al.*, 2022]. However, these methods heavily depend on the fidelity of the multi-view images and often struggle to generate high-quality meshes, particularly in capturing intricate geometric details. To mitigate this issue, a newer paradigm [Zhang *et al.*, 2023a] leverages 3D variational auto-encoders to compress high-resolution point clouds into a compact latent space before performing diffusion to directly generate 3D shapes. Despite its potential, this approach is limited by the lengthy training time and the lack of a scaling strategy, which constrain its effectiveness and scalability.

In parallel, the emergence of autoregressive (AR) Large Language Models [Brown *et al.*, 2020] and multimodal AR models [Liu *et al.*, 2023] has opened a new era in artificial intelligence. These models demonstrate exceptional scalability, versatility, as well as generalization and multimodal capabilities. At the core of these AR models is the tokenizer [Esser *et al.*, 2021], which transforms diverse data into discrete tokens, enabling the model to employ self-supervised generative learning for next-token prediction.

AR models have also made notable advancements in visual generation, leveraging their sequential processing capabilities to construct images as raster-scan token grids [Yu *et al.*, 2023]. However, given the unordered and unstructured nature of 3D data, extending next-token prediciton to 3D generation tasks remains a hurdle. For example, MeshGPT [Siddiqui *et al.*, 2024] and its successors tokenize serialized mesh data with a GNN-based encoder [Zhou *et al.*, 2018] and rely on manually defined sequence ordering [Wu *et al.*, 2024b]. Despite the promising results on shapes with simple topology, they generally fail to represent complex geometry. On the other hand, recent attempts that encapsulate 3D shapes as structured 3D volumes [Cheng *et al.*, 2023; Xiang *et al.*, 2024] or 2D triplanes [Wu *et al.*, 2024a] also struggle to learn effective feature representations from unordered 3D data, due to the lack of compression expressivity, robustness, and scalability in representing high-quality geometry.

Remarkably, we argue that 3D data inherently exhibits level-of-detail characteristics, with a natural sequential relationship across different scales – a concept well established in 3D rendering [Lindstrom *et al.*, 1996] and reconstruction [Zhang *et al.*, 2021]. In light of this insight, we introduce G3PT – a scalable, coarse-to-fine 3D native generative model that effectively maps unordered point-based 3D data

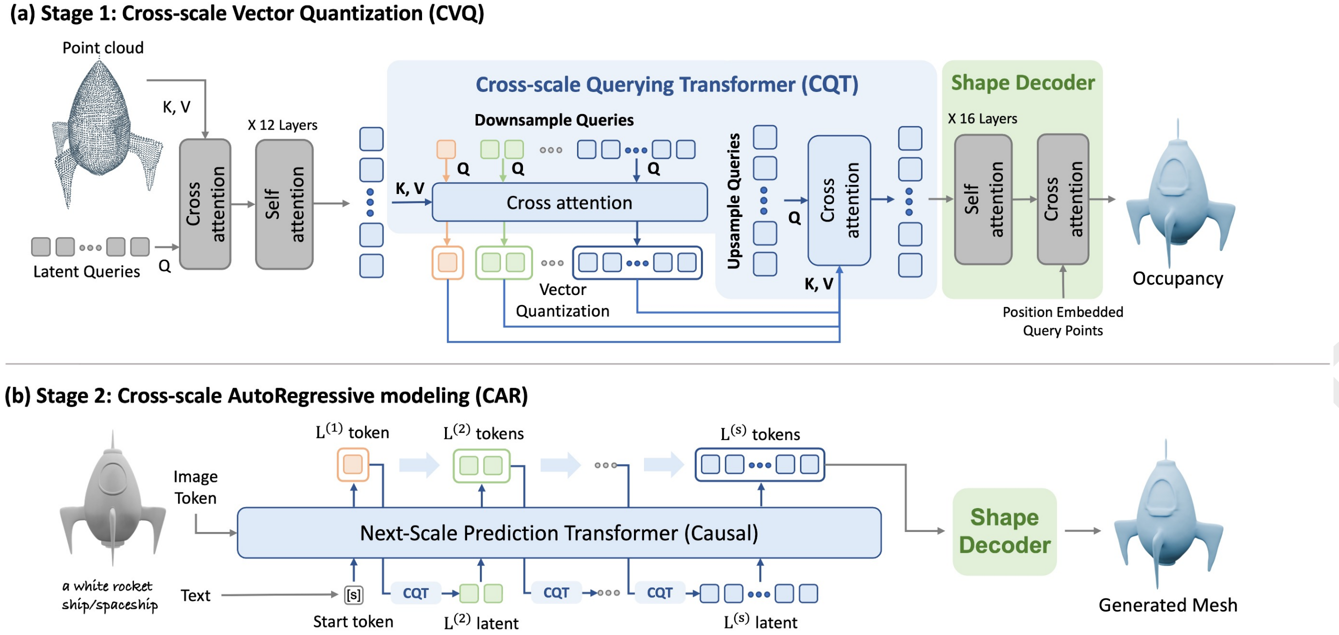

Figure 1: Overall pipeline of G3PT for representing and generating unordered 3D data. (a) Cross-scale Vector Quantization (CVQ): G3PT encodes the input point cloud into discrete scales of token vectors, each representing a different level of detail. The proposed Cross-scale Querying Transformer (CQT) utilizes cross-attention mechanisms with learnable queries of varying lengths to globally connect tokens across different scales, without requiring the tokens to be organized in a specific order. The final output is the occupancy value for each query point. (b) Cross-scale AutoRegressive modeling (CAR): G3PT reuses the CQT from the stage of CVQ for cross-scale dimension alignment and enables scalable 3D native generation from coarse to fine scales under various conditions, with an autoregressive transformer trained using next-scale prediction.

into discrete tokens at various levels of detail, creating a sequential relationship ideally suited for autoregressive modeling. Unlike the recent Visual AutoRegressive [Tian *et al.*, 2024a] (VAR) model, which also use "next-scale prediction" but rely on average pooling and bilinear interpolation (operations poorly suited for unordered data), G3PT employs Cross-scale Querying Transformer (CQT), which uses cross-attention and learnable queries to connect tokens across different scales, to effectively enable global integration of information without imposing a specific token order.

The training of G3PT comprises two stages, namely Cross-scale Vector Quantization (CVQ) and Cross-scale AutoRegressive modeling (CAR). In CVQ, we employ a transformer-based tokenizer to encode high-resolution point clouds into latent tokens and decode them into 3D occupancy grids through querying points [Zhang *et al.*, 2023a]. During this process, CQT is utilized to decompose the latent feature into discrete tokens at various scales, yielding a level-of-detail 3D representation for residual vector quantization [Tian *et al.*, 2024a] (Figure 1 (a)). With the cross-scale quantized tokens, the CAR process of G3PT begins at the coarsest scale with only one token, and the transformer predicts the next-scale token map conditioned on all previous ones by reusing the CQT for dimension alignment (Figure 1 (b)). This CAR approach provides G3PT with a versatile 3D generation pipeline, which seamlessly supports diverse conditional modalities, including image-based and text-based inputs. Extensive experiments show that G3PT not only surpasses previous LRMs and diffusion-based 3D generation methods in terms of generation quality, but also, for the first time in 3D generation, reveals distinct scaling-law behaviors.

In summary, the key contributions of this work are:

- The introduction of the first cross-scale autoregressive modeling framework for generating unordered data, offering insights to AR models on the 3D generation task.
- The development of a Cross-scale Querying Transformer (CQT) that tokenizes 3D data into discrete tokens at varying scales, enabling sequential coarse-to-fine AR modeling.
- Demonstration through extensive experiments that G3PT sets a new state-of-the-art in 3D content creation, outperforming previous LRMs and diffusion-based methods.

2 Related Work

Large reconstruction models. Extensive large-scale 3D datasets [Deitke *et al.*, 2023; Deitke *et al.*, 2024] have enabled the development of LRMs [Hong *et al.*, 2023; Tochilkin *et al.*, 2024], which utilize transformers to map image tokens to triplanes with multi-view supervision. Instant3D [Li *et al.*, 2023] and MeshLRM [Wei *et al.*, 2024] extend LRM from single-view to sparse multi-view inputs by integrating a multi-view diffusion model. Methods like InstantMesh [Xu

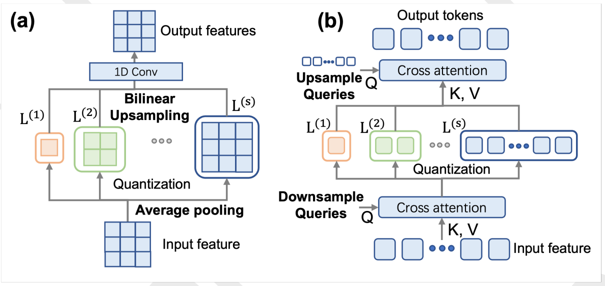

Figure 2: Comparison of multi-scale vector quantization approaches. (a) The quantization approach used in VAR [Tian *et al.*, 2024a] relies on average pooling and bilinear upsampling, which are not suitable for unordered data. (b) Our Cross-scale Vector Quantization (CVQ) overcomes this limitation with CQT, which employs a set of cross-scale learnable queries to globally "downsample" and "upsample" the unordered input feature with cross-attention. Specifically, a set of "downsample" learnable queries "pool" the input feature into token vectors of decreased length at each scale, effectively forming a level-of-detail tokenization. These cross-scale token vectors are then "upsampled" to the original scale for residual quantization, with another set of "upsample" learnable queries.

et al., 2024a] and CRM [Wang *et al.*, 2024] incorporate Flexicubes [Shen *et al.*, 2023] for direct mesh optimization. To improve rendering efficiency, LGM [Tang *et al.*, 2024] and GRM [Xu *et al.*, 2024b] replace NeRF with 3D Gaussians [Kerbl *et al.*, 2023]. However, these approaches often prioritize minimizing rendering loss over explicit mesh generation, resulting in noisy and coarse geometry.

3D native generative models. Generating 3D content with direct 3D supervision offers a more efficient approach, yet training 3D generative models directly on 3D data poses significant challenges due to high memory and computational demands. Recent methods, such as MeshGPT [Siddiqui *et al.*, 2024], Shap-E [Jun and Nichol, 2023], and others [Zhang *et al.*, 2023a; Zhao *et al.*, 2024; Li *et al.*, 2024; Zhang *et al.*, 2024], compress 3D shapes into a compact latent space before performing diffusion or autoregressive processes. While MeshGPT shows promise, its performance is limited by the mesh tokenizer. Direct3D [Wu *et al.*, 2024a] and LAM3D [Cui *et al.*, 2024] further enhance generation quality by introducing explicit 3D triplane representations. Make-A-Shape [Hui *et al.*, 2024] and WaLa [Sanghi *et al.*, 2024] use a wavelet-tree representation to enhance geometry encoding. TRELLIS [Xiang *et al.*, 2024] is a very recent, open-source 3D native generative model, whose latent representation is based on the hierarchical and sparse 3D voxels. It is composed of a sparse 3D VAE and a rectified flow transformer, enabling the generation of versatile 3D representations. However, the extended training cycles and unpredictable scaling behaviors still constrain the efficiency of these 3D generation approaches.

Autoregressive models for image generation. Autoregressive models have revolutionized visual generation by sequentially creating images using discrete tokens, produced by image tokenizers [Van Den Oord *et al.*, 2017; Esser *et al.*, 2021]. Models like DALL-E [Ramesh *et al.*, 2021], RQ-Transformer [Lee *et al.*, 2022], and Parti [Yu *et al.*, 2022] rely on raster-scan sequences for "next token prediction" within a given scale. VAR [Tian *et al.*, 2024b] introduces a novel "next-scale prediction" approach, which better preserves spatial locality and reduces computational costs. In this paper, we aim to explore the scalability potential of next-scale autoregressive modeling in 3D generation.

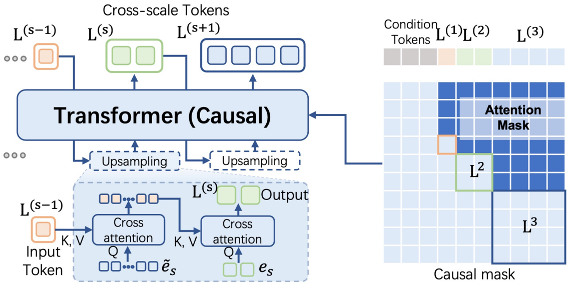

Figure 3: Illustration of the proposed CAR process in G3PT. The transformer predicts the next-scale token vector using features derived from the "upsampled" tokens of the previous scale. The "upsampling" process involves two layers of cross-attention to align the number of tokens across scales. First, features are "upsampled" with a learnable query $\tilde{e}_s$, and then "downsampled" using a "downsample" query e_s to match the token number of the next scale. A causal mask is applied to maintain the correct order and dependencies across different scales and input conditions, ensuring coherence in the model predictions.

3 G3PT

We introduce G3PT, a scalable and hierarchical 3D native generative model comprising two training stages, namely Cross-scale Vector Quantization (CVQ) and Cross-scale AutoRegressive Modeling (CAR). In the CVQ stage, we perform multi-scale residual quantization [Tian *et al.*, 2024a], converting unordered, point-based 3D data into compact, discrete tokens across multiple scales. The core of this process is the Cross-scale Querying Transformer (CQT), which utilizes cross-attention mechanisms and learnable ”downsample” and ”upsample” queries to efficiently generate a level-of-detail latent 3D representation. In the subsequent CAR stage, this latent hierarchy is aligned by CQT and transformed into a sequential formulation for next-scale prediction.

In the following, we first review the key components of tokenization and autoregressive modeling in Section 3.1, followed by a detailed description regarding the Cross-scale Vector Quantization (CVQ) (Section 3.2) and the Cross-scale AutoRegressive Modeling (CAR) (Section 3.3).

3.1 Preliminaries

Tokenization. We use Lookup-Free Quantization (LFQ) [Yu *et al.*, 2023] to tokenize the feature map $Z \in \mathbb{R}^{L \times C}$ with L tokens and C-dimensional embeddings into the quantized feature map $\hat{Z}$. LFQ streamlines the quantization process by eliminating the need for explicit codebook lookups, thereby reducing the embedding dimension of the feature Z. Formally, the quantization is executed via a mapping function

$\zeta = q(z) = \text{sign}(z)$, which maps a feature vector $z \in \mathbb{R}^C$ to an index vector $\zeta \in \mathbb{R}^{\log_2 C}$, with each dimension of ζ being quantized independently. The token index for $q(z)$ using LFQ is determined by:

$$\text{Index}(z) = \sum_{i=1}^{\log_2 C} 2^{i-1}\mathbb{1}\{z_i > 0\}. \tag{1}$$

Autoregressive modeling. For a sequence of discrete tokens $x = (x_1, x_2, \ldots, x_N)$, the probability distribution over the sequence is defined as the product of the conditional probabilities of each token given its predecessors, expressed as $P(x) = \prod_{i=1}^{N} P(x_i \mid x_1, x_2, \ldots, x_{i-1})$. This approach effectively models the dependencies between tokens, which is crucial in generating coherent sequences in tasks like language processing and image synthesis.

To account for the spatial dependencies and multi-scale characteristics of images, next-scale prediction [Tian *et al.*, 2024a] is proposed to progressively refine the 2D latent representation across a sequence of varying scales. However, the original implementation constructs multi-scale token maps using average pooling and bilinear interpolation, both of which assume an inherent order on the tokens at each scale. This ordering becomes problematic when applied to unordered 3D data.

3.2 Cross-Scale Vector Quantization (CVQ)

Shape encoding. As illustrated in Figure 1 (a), we first follow the architecture described in 3DShape2VecSet [Zhang *et al.*, 2023b] to encode 3D shapes. The input point cloud is represented as $X \in \mathbb{R}^{N\times(3+3)}$, with each of the N points having 3 position and 3 normal point features. We employ a cross-attention layer to integrate the 3D information from X into the learnable latent queries $Lat \in \mathbb{R}^{L\times C}$, as follows:

$$Z = \text{CrossAttn}(Lat, \text{PosEmb}(X)), \tag{2}$$

where PosEmb represents Fourier positional encoding and $Z \in \mathbb{R}^{L\times C}$ is the output latent features.

CVQ. We then perform Cross-scale Vector Quantization (CVQ), a novel quantization process that extends the residual quantization steps [Tian *et al.*, 2024a] leveraging the Cross-scale Querying Transformer (CQT).

A high-level comparison between the quantization used in VAR and the proposed CVQ is illustrated in Figure 2. As shown in Figure 2 (a), the conventional downsampling and upsampling operations, such as average pooling and bilinear interpolation, require a sequential arrangement of tokens, which is problematic for unordered tokens. To address this issue, we propose Cross-scale Querying Transformer (CQT) to yield a level-of-detail latent 3D representation using a set of “downsample” and “upsample” learnable queries and cross-attention, as shown in Figure 2 (b).

Specifically, we first apply a cross-attention layer and introduce a set of cross-scale learnable queries $\{e_1, e_2, \ldots, e_S\}$ – with a hierarchy of decreased token numbers – to “downsample” the unordered 3D tokens Z. More formally, at each scale s, the current residual feature $Z_s \in \mathbb{R}^{L\times C}$ (initialized by $Z_S = Z$ and $L^{(S)} = L$), is “downsampled” into a compact latent vector $E_s \in \mathbb{R}^{L^{(s)}\times C}$ by a cross-attention layer with a “downsample” learnable query $e_s \in \mathbb{R}^{L^{(s)}\times C}$:

$$E_s = \text{CrossAttn}_{down}(e_s, Z_s). \tag{3}$$

In this configuration, e_s serves as the query head, while Z_s acts as the key and value heads. Following Eq. (1), LFQ is applied on E_s, and we denote by $\hat{E}_s$ the resulting quantized token vector of the latent feature E_s.

Then, we retrieve the “upsampled” feature $\tilde{Z}_s \in \mathbb{R}^{L\times C}$ of the original scale by using another cross-attention layer with an “upsample” learnable query $\tilde{e}_s \in \mathbb{R}^{L\times C}$:

$$\tilde{Z}_s = \text{CrossAttn}_{up}(\tilde{e}_s, \hat{E}_s). \tag{4}$$

The residual feature for the subsequent scale, Z_{s+1}, is then calculated as: $Z_{s+1} = Z_s - \tilde{Z}_s$. This process iteratively continues until the final quantization step S.

Shape decoding. The final 3D shape decoder, as shown on the right side of Figure 1 (a), which consists of several self-attention layers and a cross-attention layer, decodes the sum of these “upsampled” features and a set of query points p into occupancy values:

$$Occ(p) = \text{CrossAttn}(\text{PosEmb}(p), \text{SelfAttn}(\sum_{s=1}^{S} \tilde{Z}_s)). \tag{5}$$

3.3 Cross-Scale AutoRegressive Modeling (CAR)

After the stage of CVQ, we obtain a sequence of cross-scale 3D token vectors $(\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_S)$ of varying lengths $(L^{(1)}, L^{(2)}, \ldots, L^{(S)})$ to serve as inputs for AR modeling. To enable effective next-scale prediction, we introduce Cross-scale AutoRegressive modeling (CAR), which reuses the CQT trained during the CVQ stage to align the token dimensions across different scales. Specifically, we first employ the “upsample” queries $\tilde{e}_{s-1} \in \mathbb{R}^{L\times C}$ to elevate the token length $L^{(s-1)}$ to the original scale L, followed by using “downsample” queries $e_s \in \mathbb{R}^{L^{(s)}\times C}$ that compress them to the scale $L^{(s)}$ to be predicted next.

To perform next-scale prediction [Tian *et al.*, 2024a], we model the probability distribution $P(\hat{E})$ over the sequence of the cross-scale 3D token vectors $(\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_S)$, where the token vector at each scale $\hat{E}_s$ is conditioned on that at the preceding coarser scales:

$$P(\hat{E}) = \prod_{s=1}^{S} P(\hat{E}_s \mid \hat{E}_1, \hat{E}_2, \ldots, \hat{E}_{s-1}). \tag{6}$$

In this manner, the token vector at each scale is added with more detailed information based on the previous one, allowing the model to progressively refine the data from a rough approximation to a detailed representation.

3.4 Conditional Autoregressive Modeling

Image conditioning. In each CAR block, pixel-level information from the conditional image and point-level information from the point cloud are seamlessly integrated, so as to align the image feature space with the 3D latent space for

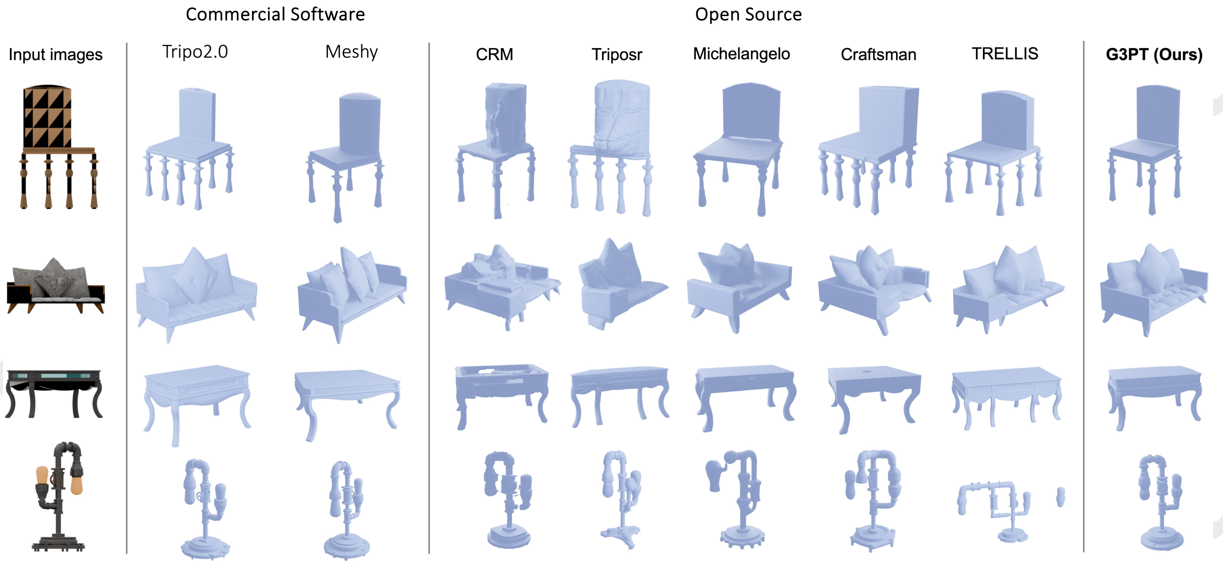

Figure 4: Qualitative comparisons on image-to-3D shape generation against state-of-the-art methods on the Objaverse dataset.

high-quality controllable 3D generation. The overall framework of the conditional CAR model is depicted in Figure 1 (b). Specifically, we employ the pre-trained DINO-v2 (ViT-L/14) [Oquab *et al.*, 2023] to extract image features, leveraging its strength in capturing the structural information crucial for 3D tasks. A linear layer projects the L_I image tokens $I_{dino} \in \mathbb{R}^{L_I \times C_I}$, derived from DINO-v2, to $Z_{dino} \in \mathbb{R}^{L_I \times C}$, which matches the channel dimension of the cross-scale 3D tokens $(\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_S)$. These projected image tokens are then concatenated with the cross-scale 3D tokens and regulated through an attention mask in the causal transformer, as shown in Figure 3, to ensure that only subsequent 3D tokens are predicted.

Text conditioning. Text conditions can also be easily added to ensure semantic consistency using the pre-trained CLIP model, which extracts semantic tokens from the conditional text input. We use AdaLN [Wu *et al.*, 2024a] for effective signal control. Note that this approach only serves as a preliminary attempt to validate the flexible conditioning capabilities of the proposed G3PT, while other advanced conditioning mechanisms are left for further exploration.

4 Experiments

4.1 Implementation Details

CVQ. Each input point cloud to CVQ contains 16384 points uniformly sampled from the 3D model in the Objaverse dataset [Deitke *et al.*, 2023], accompanied by a learnable latent query with length $L = 2304$ and channel dimension $C = 512$. The vocabulary size of the codebook in LFQ is 8192. The shape encoder network includes one cross-attention layer and 12 self-attention layers. The shape decoder network contains a cross-attention layer and 16 self-attention layers with the same channel dimension of the encoder. When training CVQ, 8192 uniform points and 8192 near-surface points are sampled for supervision. The AdamW optimizer is employed with a learning rate of 1×10^{-4}, and the CVQ model is trained for 60,000 steps on 8 NVIDIA A100 GPUs with 80GB memory.

CAR. The transformer in CAR shares a similar architecture with the standard decoder-only transformer used in GPT-2. To stabilize training, queries and keys are normalized to unit vectors before calculating the attention weights. All models are trained with a learning rate of 1×10^{-5} and a batch size of 1600, using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.05 for every 1000 steps. The 1.5B model is trained for two weeks on 136 NVIDIA H20 GPUs with 96GB memory.

Type	Method	Name	IoU↑	Cham.↓	F-score↑
LRM	NeRF	Triposr	72.6	0.023	58.2
		InstantMesh	68.7	0.029	58.3
		CRM	76.3	0.020	61.4
	Gaussian	LGM	67.6	0.025	49.3
3D Generation	Diffusion	Michelangelo	74.5	0.028	62.5
		Shap-E	66.8	0.029	46.3
		CraftsMan	72.2	0.021	56.1
		Make-A-Shape	69.3	0.025	54.9
		WaLa	71.4	0.025	62.6
		CLAY*	77.1	0.021	63.4
		TRELLIS	79.9	0.021	70.4
	AR Modeling	G3PT (0.1B)	73.9	0.025	60.4
		G3PT (0.5B)	82.11	0.015	75.1
		G3PT (1.5B)	**87.6**	**0.013**	**83.0**

Table 1: Quantitative comparisons on image-to-3D shape generation against state-of-the-art LRMs and diffusion-based 3D generation approaches on the Objaverse dataset. (*Reproduction)

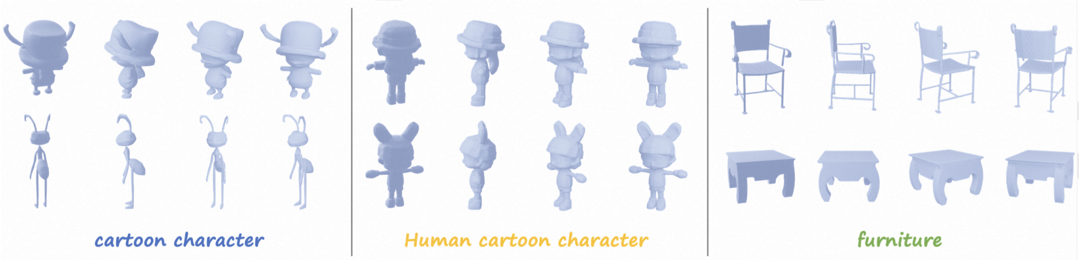

Figure 5: Text-to-3D generation results using G3PT. The colored texts serve as prompt conditions for text-to-3D shape generation.

Method	#Token	IOU↑	Cham.↓	F-score↑	Acc.(%)↑	Usage(%)↑
VAE	576	89.20	0.0126	84.10	95.24	-
	2408	89.60	0.0118	85.80	95.80	-
VQVAE	576	85.32	0.0134	80.15	85.59	96.34
	2408	87.43	0.0131	80.53	88.32	92.96
CVQ	576	89.38	0.0122	85.70	95.27	**99.51**
	2408	**90.35**	**0.0108**	**87.23**	**97.13**	97.13

Table 2: Ablative comparisons on different tokenizers.

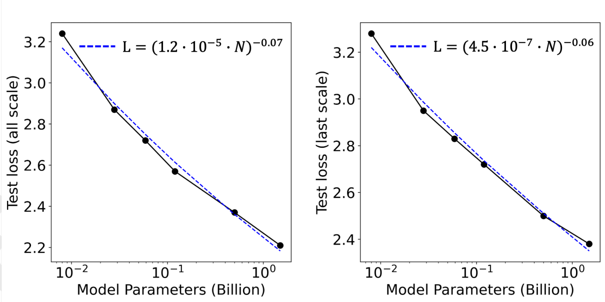

Figure 6: Scaling laws in G3PT with network parameters N.

4.2 Training Details

CVQ. Directly training with a large number of discrete tokens is highly time-consuming. To mitigate this, during the initial training phase for CVQ, the quantization layer in between the encoder and decoder is replaced by a layer normalization to facilitate convergence. Once this model is adequately trained, the CVQ is further finetuned with the quantization layer.

CAR. We implement a progressive training strategy for CAR. Specifically, instead of processing the tokens across all scales $(L^{(1)}, L^{(2)}, \dots, L^{(S)})$ at once, the training begins with tokens before the $S/2$ scale $(L^{(1)}, L^{(2)}, \dots, L^{(S/2)})$ and progressively includes finer scales. This approach accelerates convergence and improves training stability.

4.3 3D Generation Results

Evaluation protocols. We mainly benchmark on the task of image-to-3D shape generation, where a single RGB image is used as the conditional signal and the output is the generated 3D mesh. We evaluate the mesh quality using Intersection-over-Union (IoU), Chamfer distance (Cham.), and F-score (with a threshold of 0.01), which reflect the overall proximity from the generated mesh to the ground-truth. The experiment is conducted on 120 randomly selected testing objects from the Objaverse dataset [Deitke *et al.*, 2023]. We compare against LRMs including InstantMesh [Xu *et al.*, 2024a], CRM [Wang *et al.*, 2024], Triposr [Tochilkin *et al.*, 2024], and LGM [Tang *et al.*, 2024], as well as diffusion-based approaches like Michelangelo [Zhao *et al.*, 2024], Shap-E [Jun and Nichol, 2023], CraftsMan [Li *et al.*, 2024], Make-A-Shape [Hui *et al.*, 2024], WaLa [Sanghi *et al.*, 2024], CLAY [Zhang *et al.*, 2024], and TRELLIS [Xiang *et al.*, 2024].

Image-to-3D generation. The results in Table 1 highlight the superiority of G3PT for image-to-3D shape generation, particularly the model with 1.5 billion parameters, which outperforms all other methods with a substantial margin in all metrics, demonstrating the unparalleled generation quality and fidelity. We also conduct qualitative evaluations by comparing the proposed G3PT with other state-of-the-art methods (including commercial software like Meshy[1] and Tripo2.0[2]) using the Objaverse dataset [Deitke *et al.*, 2023]. As shown in Figure 4, LRMs like CRM and Triposr suffer from noisy artifacts. Diffusion-based methods like Michelangelo and TRELLIS produce plausible geometry but struggle to align with the semantic structure of the conditional images. In contrast, G3PT achieves a superior balance between geometry quality and controllability, consistently producing high-quality meshes that align well with the conditional images. Note that the proposed G3PT, only trained on the Objaverse dataset, performs on par with Meshy and Tripo2.0, which span a large amount of internal data for training. Please refer to the supplement for more quantitative and qualitative evaluations on other datasets.

Text-to-3D generation. In Figure 5, we present the qualitative generation results of G3PT, given only text prompts as conditions. The generated meshes align well with the conditional semantics while exhibit remarkable diversity and high-quality geometry.

Scaling Behaviors. Figure 6 illustrates the scaling laws observed in G3PT by examining the relationship between the number of model parameters (in billions) and the test loss (measured as cross-entropy). Both plots demonstrate a clear

[1] https://www.meshy.ai/discover

[2] https://www.tripo3d.ai/

#Token	Codebook Size	IOU(%)↑	Cham.↓	F-score(%)↑	Acc.(%)↑
2408	1024	85.43	0.0122	81.25	92.96
2408	2048	86.65	0.0115	86.36	94.44
2408	4096	89.32	0.0119	87.46	96.21
2408	**8192**	**90.35**	**0.0108**	**87.23**	**97.13**

(a) Codebook sizes.

#Token	Codebook Size	IOU(%)↑	Cham.↓	F-score(%)↑	Acc.(%)↑
256	8192	84.31	0.0125	80.18	93.32
576	8192	89.38	0.0122	85.7	95.27
1024	8192	89.51	0.0119	86.86	96.58
2408	8192	**90.35**	**0.0108**	**87.23**	**97.13**

(b) Number of tokens.

Table 3: Ablation on different codebook sizes and number of tokens.

3D Representation	Encoding Method	Quantization	IOU (%) ↑	Cham. ↓	F-score (%) ↑	Acc. (%) ↑	Usage (%) ↑
Volume	3D CNN	LFQ	85.29	0.0114	78.66	86.89	91.34
Triplane	Learnable query	LFQ	86.13	0.0120	80.12	90.44	89.34
1D Latent	Learnable query	Pooling + LFQ	89.51	0.0139	86.86	93.80	**99.50**
	Learnable query	CVQ	**90.35**	**0.0108**	**87.23**	**97.13**	97.13

Table 4: Comparison of various 3D representations paired with different encoding methods.

trend where the test loss decreases as the number of model parameters increases, with a power-law relationship. This further validates the potential of G3PT in handling complex 3D generation tasks.

4.4 Ablation Studies

In the following, we perform comprehensive ablation studies on the task of image-to-3D shape generation, strictly following the evaluation protocols outlined in Section 4.3.

Tokenization. We first demonstrate the effectiveness of the proposed CVQ by comparing to other tokenizers. Table 2 presents the comparison against two commonly used approaches: VAE and VQVAE. We implement VAE following [Zhang *et al.*, 2024], which applies KL regularization between the encoder and decoder, differing from the quantization module utilized in diffusion models. We implement VQ-VAE by incorporating LFQ quantization, which maintains the same quantization structure as described by [Yu *et al.*, 2023]. The quantitative metrics additionally include prediction accuracy (Acc.) of the occupancy value (0 or 1), which is determined by evaluating points that are randomly sampled in the vicinity of the ground-truth mesh. The “Usage” metric indicates the efficiency of codebook usage. As can be seen, CVQ outperforms both VAE and VQVAE across multiple metrics with near-complete codebook usage.

Codebook size and the number of tokens. A detailed quantitative comparison given different codebook sizes and the number of tokens (#token) are presented in Table 3 (a) and (b), respectively. Due to memory constraints, the evaluation is limited to a maximum of 8192 tokens. We refer readers to the supplement for relationships between #token and #scale.

3D representation. The comparison of different 3D representations and encoding methods is reported in Table 4. We consider three types of 3D representations: volumetric feature grids (Volume), Triplane [Wang *et al.*, 2023], and 1D latent vectors. The Volume representation is encoded using a 3D Convolutional Neural Network (3D CNN) following the architecture of SDFusion [Cheng *et al.*, 2023], while the Triplane representation utilizes the architecture from Direct3D [Wu *et al.*, 2024a]. All encoders are designed with similar parameter counts. For 1D Latent, a comparison is also conducted between CVQ and a baseline setup (Pooling + LFQ), i.e., a similar implementation to VAR [Tian *et al.*, 2024a] by incorporating a 1D average pooling module, a bilinear upsampling module, and an additional 1D convolutional layer (Figure 2 (a)), which forces the tokens to learn an ordered sequence.

The results demonstrate that CVQ achieves superior performance with near-complete codebook usage, emphasizing the effectiveness of CVQ in preserving detailed structural information during quantization. When compared to the baseline setup with average pooling and bilinear upsampling, the results are in line with our intuition that 1D latent tokens do not possess an inherent sequential order, unlike image data.

5 Conclusion

This paper introduces G3PT, a scalable hierarchical 3D generative model featuring a Cross-scale Querying Transformer (CQT) to map unordered 3D data into discrete tokens across various levels of detail. By establishing a natural sequential relationship among these tokens, G3PT enables Cross-scale AutoRegressive modeling (CAR) in a manner that aligns well with the inherent unordered characteristics of 3D data. This novel CAR framework tailored for unordered data offers new insights into autoregressive algorithm design. Extensive experiments demonstrated that G3PT achieves superior generation quality compared to existing 3D generation methods, setting a new state-of-the-art in 3D content creation.

References

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Cheng *et al.*, 2023] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tuyakov, Alex Schwing, and Liangyan Gui. SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In *CVPR*, 2023.

[Cui *et al.*, 2024] Ruikai Cui, Xibin Song, Weixuan Sun, Senbo Wang, Weizhe Liu, Shenzhou Chen, Taizhang Shang, Yang Li, Nick Barnes, Hongdong Li, et al. Lam3d: Large image-point-cloud alignment model for 3d reconstruction from single image. *arXiv preprint arXiv:2405.15622*, 2024.

[Deitke *et al.*, 2023] 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*, pages 13142–13153, 2023.

[Deitke *et al.*, 2024] 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.

[Esser *et al.*, 2021] 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.

[Hong *et al.*, 2023] 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.

[Hui *et al.*, 2024] Ka-Hei Hui, Aditya Sanghi, Arianna Rampini, Kamal Rahimi Malekshan, Zhengzhe Liu, Hooman Shayani, and Chi-Wing Fu. Make-a-shape: a ten-million-scale 3d shape model. In *Forty-first International Conference on Machine Learning*, 2024.

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

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

[Lee *et al.*, 2022] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[Li *et al.*, 2023] 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.

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

[Lindstrom *et al.*, 1996] Peter Lindstrom, David Koller, William Ribarsky, Larry F Hodges, Nick Faust, and Gregory A Turner. Real-time, continuous level of detail rendering of height fields. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 109–118, 1996.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[Oquab *et al.*, 2023] 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.

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

[Ramesh *et al.*, 2021] 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.

[Rombach *et al.*, 2022] 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.

[Sanghi *et al.*, 2024] Aditya Sanghi, Aliasghar Khani, Pradyumna Reddy, Arianna Rampini, Derek Cheung, Kamal Rahimi Malekshan, Kanika Madan, and Hooman Shayani. Wavelet latent diffusion (wala): Billion-parameter 3d generative model with compact wavelet encodings, 2024.

[Shen *et al.*, 2023] 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):37–1, 2023.

[Siddiqui *et al.*, 2024] 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.

[Tang *et al.*, 2024] 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.

[Tian *et al.*, 2024a] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

[Tian *et al.*, 2024b] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

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

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[Wang *et al.*, 2023] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. Pet-neus: Positional encoding tri-planes for neural surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12598–12607, 2023.

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

[Wei *et al.*, 2024] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024.

[Wu *et al.*, 2024a] 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.

[Wu *et al.*, 2024b] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024.

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

[Xu *et al.*, 2024a] 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.

[Xu *et al.*, 2024b] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024.

[Yu *et al.*, 2022] 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.

[Yu *et al.*, 2023] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[Zhang *et al.*, 2021] Jinzhi Zhang, Mengqi Ji, Guangyu Wang, Zhiwei Xue, Shengjin Wang, and Lu Fang. Surrf: Unsupervised multi-view stereopsis by learning surface radiance field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7912–7927, 2021.

[Zhang *et al.*, 2023a] 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.

[Zhang *et al.*, 2023b] 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), jul 2023.

[Zhang *et al.*, 2024] 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.

[Zhao *et al.*, 2024] 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.

[Zhou *et al.*, 2018] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *ArXiv*, abs/1812.08434, 2018.