Auto-Regressively Generating Multi-View Consistent Images

JiaKui Hu[1,2,3,4*], Yuxiao Yang[5,2*], Jialun Liu[2†], Jinbo Wu[2‡], Chen Zhao[2], Yanye Lu[1,3,4†]
[1]Institute of Medical Technology, Peking University Health Science Center, Peking University
[2]Baidu VIS
[3]Biomedical Engineering Department, College of Future Technology, Peking University
[4]National Biomedical Imaging Center, Peking University
[5]Tsinghua University
jkhu29@stu.pku.edu.cn, yangyuxi23@mails.tsinghua.edu.cn, liujialun95@gmail.com,
yanye.lu@pku.edu.cn

Abstract

*Generating multi-view images from human instructions is crucial for 3D content creation. The primary challenges involve maintaining consistency across multiple views and effectively synthesizing shapes and textures under diverse conditions. In this paper, we propose the Multi-View Auto-Regressive (**MV-AR**) method, which leverages an auto-regressive model to progressively generate consistent multi-view images from arbitrary prompts. Firstly, the next-token-prediction capability of the AR model significantly enhances its effectiveness in facilitating progressive multi-view synthesis. When generating widely-separated views, MV-AR can utilize all its preceding views to extract effective reference information. Subsequently, we propose a unified model that accommodates various prompts via architecture designing and training strategies. To address multiple conditions, we introduce condition injection modules for text, camera pose, image, and shape. To manage multi-modal conditions simultaneously, a progressive training strategy is employed. This strategy initially adopts the text-to-multi-view (t2mv) model as a baseline to enhance the development of a comprehensive X-to-multi-view (X2mv) model through the randomly dropping and combining conditions. Finally, to alleviate the overfitting problem caused by limited high-quality data, we propose the "Shuffle View" data augmentation technique, thus significantly expanding the training data by several magnitudes. Experiments demonstrate the performance and versatility of our MV-AR, which consistently generates consistent multi-view images across a range of conditions and performs on par with leading diffusion-based multi-view image generation models. The code and models are released here.*

1. Introduction

Generating multi-view images from human instructions, *e.g.*, texts, reference images, and geometric shapes, represents a crucial endeavor with profound implications in domains such as 3D content creation, robotic perception, and simulation. The challenge lies in the development of models capable of generating multi-view consistent images while addressing diverse prompts in a unified architecture.

Recent multi-view image generation approaches [17, 21–23, 35, 36] concentrate on leveraging the image generation priors of the pre-trained diffusion model [33]. With the advancement of video diffusion models, some methods have evolved to utilize pre-trained video diffusion models [3] for multi-view images with exemplary outcomes [48]. As shown in Figure 1, these methods utilize text or images as conditions, facilitating simultaneous synthesis across multiple views. However, when synthesizing images from distant views, the overlap between reference and target images decreases significantly, weakening the effectiveness of reference guidance. In extreme cases, such as generating a back-view image from a front-view reference, the visual reference information becomes nearly negligible due to minimal overlapping textures. Consequently, this insufficient reference between distant views severely compromises the multi-view consistency.

To address this limitation, we propose employing Auto-Regressive (AR) generation for multi-view image generation. As shown in Figure 1, in AR-based generation, the model leverages information derived from the preceding $n-1$ views as conditions to generate the n-th view, allowing the model to utilize information derived from previously generated views. This generation process is highly consis-

*This work was done when they interned in Baidu. Equal contribution.
†Corresponding author.
‡Project leader.

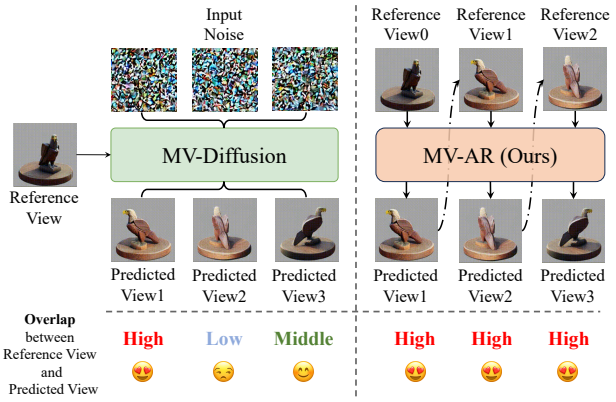

Figure 1. Diffusion-based multi-view image generation methods use a specific reference view for predicting subsequent views, which becomes problematic when overlap between the reference view and the predicted view is minimal, affecting image quality and multi-view consistency. Our MV-AR addresses this by using the preceding view with significant overlap for conditioning.

tent with how humans observe 3D objects. For instance, in scenarios where a back-view is generated from a front-view reference, the AR generation model extracts adequate and pertinent reference from the preceding views. Inspired by this concept, we propose the Multi-View Auto-Regressive (**MV-AR**) model. MV-AR establishes a novel framework for generating multi-view consistent images, which uses all previous view information to predict patches of subsequent views, thus facilitating the model's ability to capture efficient reference for distant views.

However, new framework introduces new practical issues. This study is dedicated to the examination of the issues [1] associated with the application of AR models in the context of multi-view image generation and proposes solutions thereto. We delineate three main issues that inhibit the effective development of AR models within multi-view image generation: *insufficient conditions*, *limited high-quality data*, and *cumulative error*. We systematically address these problems from two distinct perspectives: *architectures* and *training*. We devise effective methods for conditional injection that apply to text [32, 36], camera pose [15], reference image [21, 22], and shape [32, 44] in our MV-AR. Specifically, we practically demonstrate that the in-context image condition is not suitable for the current AR model. We design a precise Image Warp Controller (IWC) module that extracts features of the overlap between the current view and the previous views. IWC inserts condition token-by-token, thereby ensuring accurate image control. Next, we progressively drop or combine multi-modal conditions, so that the model can handle multi-modal conditions simultaneously after training. Finally, to address the challenge of limited training data, we formulate a data augmentation termed "Shuffle Views" and a progressive learning strategy to enrich the training data and mitigate the risk of overfitting. Utilizing the aforementioned solutions, we formulate a series of guidelines for the AR Model in multi-view image generation and establish a new baseline for its application.

In summary, our contributions to the multi-view image generation field include:

- **Auto-Regressively generating multi-view images**. We first introduce the Auto-Regressive (AR) framework in the multi-view image generation task. AR can effectively capture the information derived from all the preceding views as conditions, thereby improving multi-view consistency across far-away views.
- **Architectures and training design**. We investigate the unique challenges faced by AR models in multi-view image generation and implement targeted designing with respect to both architecture and training strategies. Based on this, we establish a robust baseline for AR-based multi-view generation approaches.
- **Performance and versatility.** We significantly narrow the performance gap between AR-based and diffusion-based multi-view generation methods. We also develop the first unified (X2mv) multi-view generation model, capable of handling various conditions such as text, image, and shape synchronously.

2. Related work

2.1. Multi-view Diffusion Models

To generate consistent multi-view images, MVDream [36] proposes to generate multi-view images conditioned on a given text prompt via a cross-view attention mechanism. Zero123++ [35] tiles multi-view images into a single frame, performs a single pass for multi-view generation, and extends the pre-trained self-attention layer to facilitate cross-view information exchange, a technique also employed in Direct2.5 [24] and Instant3D [18]. Syncdreamer [22] integrates multi-view features into 3D volumes, conducting 3D-aware fusion in 3D noise space. Wonder3D [23] introduces a multi-view cross-domain diffusion model that enhances cross-domain and cross-view information exchange through distinct attention layers. However, all of the aforementioned methods share the same core idea: modeling 3D generation using a multi-view joint probability distribution via an additional attention mechanism, which is inherently unnatural and fails to capture the progressive nature of novel view generation. Moreover, diffusion-based multi-view generation methods [19, 20, 22, 23, 35, 36] often require significant architectural modifications when the conditioning type changes. In contrast, we propose a method that generates multi-view images in an auto-regressive manner and incor-

[1] This study focuses on addressing the AR model's issues associated with generating multi-view images. The resolution of fundamental issues inherent to AR models, including their unidirectional nature and discrete encoding, is designated for future work.

porates multi-modal conditions in a unified framework.

2.2. Autoregressive Visual Generation

Pioneered by PixelCNN [41], researchers have proposed generating images as sequences of pixels. Early research, including VQVAE [42] and VQGAN [10], quantizes image patches into discrete tokens and employs transformers to learn Auto-Regressive (AR) priors, similar to language modeling [4]. To further improve reconstruction quality, RQVAE [16] introduces multi-scale quantization, while VAR [39] reformulates the process into next-scale prediction, significantly enhancing sampling speed. Parallel efforts have been dedicated to scaling up AR models for text-conditioned visual generation tasks [11]. As for the 3D area, although some preliminary works [27, 37] employ AR models to directly generate vertices, faces of meshes, they are limited to the geometry domain and struggle to scale to general 3D object datasets [8]. Unlike the aforementioned methods, our model is based on a pre-trained text-to-image AR model [38], leveraging its strong generation prior and extending it to multi-view generation tasks conditioned on diverse input types.

3. Methods

3.1. Auto Regressive Model in Multi-View

Formulation. The vanilla auto-regressive (AR) model is designed to infer the distribution of a long sequence. Specifically, given a sequence x of length T, the AR model seeks to derive the distribution of the x according to the following formula:

$$p(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_{<t}), \tag{1}$$

where x_i denotes the i-th datapoint in sequence x and $x_{<t} = (x_1, x_2, \ldots, x_{t-1})$ denotes the vector of random variables with index less than t. Training an AR model p_θ involves optimizing $p_\theta(x_t|x_{<t})$ over large-scale image sequences.

In image generation tasks, these sequences are generated by vision tokenizers [10, 16, 42]. The encoder $\varepsilon(\cdot)$ extracts the high-dimensional features $f \in \mathbb{R}^{h\times w\times D}$ of an image I. Subsequently, f is flattened into discrete 1d tokens $q \in [V]^{h\times w}$. The quantizer $\mathcal{Q}$ generally comprises a downloadable codebook $Z \in \mathbb{R}^{V\times C}$ that contains C vectors. In the quantization process $q = \mathcal{Q}(f)$, each feature vector $f(i,j)$ is assigned to the code index $q^{(i-1)*w+j}$ of its closest entry in the codebook:

$$q^{(i-1)*w+j} = \left(\arg\min_{v\in[V]} \|\text{lookup}(Z, v) - f^{(i,j)}\|_2 \right) \in [V], \tag{2}$$

where the term $\text{lookup}(Z, v)$ refers to the retrieval of the v-th vector in the codebook Z. The notation indexed as superscript $(i-1)*w+j$ of q signifies that q is flattened, whereby the data originally located at the 2d coordinate (i, j) are transformed into its 1d counterpart $(i-1)*w+j$.

Reformulation. Under the multi-view image generation task, the construction of the sequences q is different from that of a single 2D image. Specifically, we use 2D VQ-VAE [38] to extract sequences of N-view images' features $f = (f_1, f_2, \cdots, f_N)$:

$$q^{(n-1)*h*w+(i-1)*w+j} = \left(\arg\min_{v\in[V]} \|\text{lookup}(Z, v) - f_n^{(i,j)}\|_2 \right) \in [V], \tag{3}$$

where n means the n-th view.

Issues of AR model in multi-view image generation. The aforementioned training sequences facilitate the model's capacity to ensure effective reference across distant views, thus generating consistent multi-view images. However, this framework introduces several issues:

- **Issue 1: Insufficient conditions**. The task of generating multi-view images necessitates that the model adeptly extract features from various conditions and generate multi-view images that maintain consistency with the given conditions. The methods for conditional injection in AR models have not been extensively studied, thereby complicating the efficient utilization of external conditions, *e.g.*, camera poses, reference image, and geometric shape.
- **Issue 2: Limited high-quality data**. AR models have been empirically shown to require a substantial volume of high-quality data (such as billions of texts [1]) to achieve saturated model training without overfitting. However, the challenges associated with the collection of 3D objects, coupled with the paucity of high-quality multi-view images, significantly hinder MV-AR training adequacy.
- **Issue 3: Cumulative error**. Within AR generation, the sequence of images from subsequent $n-1$ views serves as conditions to generate the image in the n-th view. In some cases, there exists a low quality image, labeled view m ($m < n$). It is expected to serve as a reliable conditional reference for the subsequent generation of views. However, because of the low quality of the m-th view, it fails to provide effective guidance for the target views, leading to a cumulative error in AR generation.

These practical limitations call for a rethinking of AR models in the context of multi-view image generation.

3.2. Architectures

In this subsection, we focus on solving Issue 1 from the perspective of architectures.

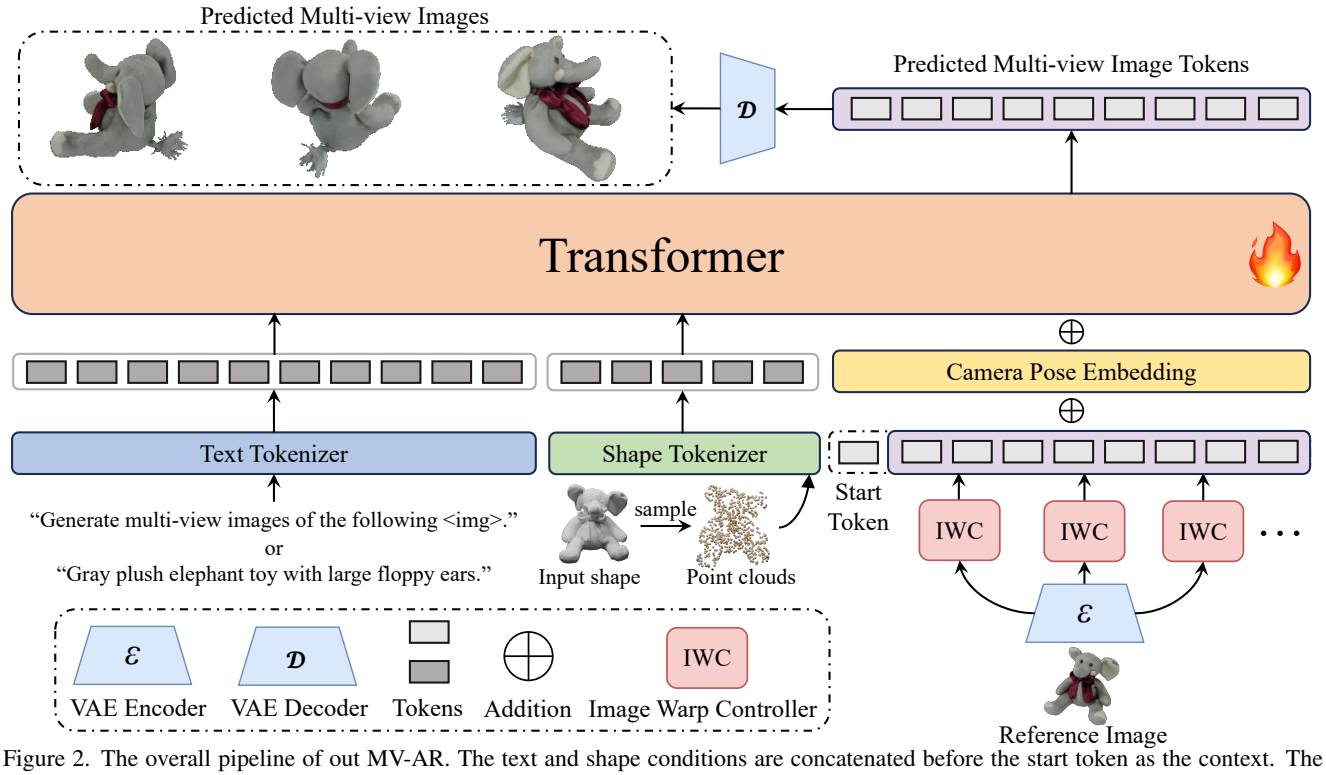

Figure 2. The overall pipeline of out MV-AR. The text and shape conditions are concatenated before the start token as the context. The text condition can either expect the model to generate multi-view images following other conditions, such as image, or describe the target object. The start token signals the model to begin generating multi-view images. Then, camera pose and image conditions are integrated. The camera pose serves as the shift position embedding, using its angular data to guide the generation of the specific view. After warping by IWC, the image conditions are added token by token within the model. It should be noted that our MV-AR can accommodate these multi-modal conditions simultaneously after progressive learning.

Transformer. Consistent with most AR models, we use Transformer [43] to learn auto-regression. Our Transformer is largely based on Llama [40], applying pre-normalization using RMSNorm [45], SwiGLU [34] activation function, and AdaLN [29].

Text Condition. To integrate the text condition into AR models, we use FLAN-T5 XL [7] as the text encoder, the encoded text feature is projected by an additional MLP and is used as prefilling token embedding in AR models.

It is crucial to note that, with this in-context text conditions, the Self-Attention (SA) mechanism in the Transformer concurrently integrates both image and text modalities. In situations where there is a misalignment between these modalities, the text tokens might be disturbed by subsequent image tokens, thereby impeding the model's capacity to be effectively directed by the text. Consequently, we develop Split Self-Attention (SSA), a mechanism that harmonizes the efficacy of SA during in-context scenarios with that of cross-attention.

$$\begin{aligned} X_{in} &= \text{Concat}(X_{text}, X_{image}); \\ O_{text}, O_{image} &= \text{Chunk}(\text{SA}(X_{in}), [N_{text}, N_{image}]); \\ \text{SSA}(X_{in}) &= X_{in} + \text{Concat}(0 \cdot O_{text}, O_{image}), \end{aligned} \tag{4}$$

where $X_{text} \in \mathbb{R}^{N_{text} \times D}$ is the text feature and $X_{image} \in \mathbb{R}^{N_{image} \times D}$ is the image feature. Concat$(\cdot)$ concatenates features in token dimension, while Chunk$(\cdot)$ is the inverse operation of Concat$(\cdot)$.

As shown in Eq. 4, SSA is achieved by forcing the output of SA at the specified condition token to zero. This approach ensures that text tokens remain unaffected by ensuing image tokens, while simultaneously maintaining the model's ability to process information pertaining to the image modality. Experiments show that SSA can improve the correspondence between the predicted multi-view images and text condition, such as the CLIP Score [12].

Camera Condition. To condition the camera pose, we use the Plücker-Ray Embedding $r \in \mathbb{R}^{(h*w)\times 6}$ that shares the same height and width as the image features f following [15]. It encodes the origin o and direction d of the ray at each spatial location.

$$r_{i,j} = (o \times d, d), \tag{5}$$

where $\times$ is the cross product.

Similarly to image features f, in our MV-AR, Plücker ray r is also required to produce a sequence as prescribed by Eq. 3. Given that r represents comprehensive angular infor-

mation and effectively conveys the positional information of tokens across various views within the entire sequence, we incorporate this ray as the Shift Position Encoding (SPE). Specifically, SPE modifies Eq. 1 to:

$$p(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} p(x_t|(x_0, x_{<t})) + r_{\leq t}), \quad (6)$$

where x_0 serves as the start token before the pre-generation image tokens as shown in Figure 2.

In Eq. 6, x_0 and $x_{<t}$ construct a sequence of length t, aligning with the sequence length of the camera condition $r_{\leq t}$, thus allowing for the direct execution of information fusion via the addition operation.

The reason for this shift lies in that r can tell the model which view or patch the next token belongs to [15]. Using r, the model can utilize this potential physical angle information to provide precise physical position data for each token within the chains of view, thus enhancing the accuracy of the model predictions for subsequent tokens.

Image Condition. Alongside the in-context conditioning aligned with the text condition, we develop the Image Warp Controller (IWC) to incorporate fine-grained image conditions into AR models. IWC uses the present camera poses r along with the features $X_{ref} = \varepsilon(I_{ref})$ from the reference view to forecast the features of the overlapped contents and textures X_{IWC} between the current and reference views, subsequently integrating them into the network in a residual fashion. IWC can be formulated as follows.

$$X_{IWC} = \text{FFN}(\text{CA}(\text{SA}(X_{ref}), r)), \quad (7)$$

where CA means the Cross Attention while FFN denotes the Feed-Forward Network.

In IWC, we do not use high-level image conditions, *e.g.*, CLIP [31] or DINO [6, 28], as an image condition. Employing low-level features that more closely approximate the intrinsic attributes of color and texture can facilitate the generation of images with detailed consistency [5, 14].

Shape Condition. The inherent one-to-many ambiguity in 3D generation often precludes precise control using text or image inputs alone (*e.g.*, an image cannot fully constrain the underlying 3D shape). However, AR models enable the injection of global shape priors as additional guidance through pre-filled token embeddings, addressing this limitation. Specifically, we adopt point clouds augmented with positional embeddings and normal maps as shape conditions, sampled from 8,192 surface points on the input mesh M. We then leverage a pre-trained shape encoder [47] to map the 3D point cloud into a fixed-length latent token sequence. To enable geometric shape control within the model, the sequence of shape tokens is strategically placed between the sequence of text tokens and the start token, as shown in Figure 2.

3.3. Training

In this subsection, we focus on solving **Issue 2** and **Issue 3** from the perspective of training. We examine approaches to address it from three critical perspectives: data augmentation and training strategy.

Loss Function. The AR model generates the conditional probability $p(q_t|q_{<t})$ of word q_t at each position t. The loss is the average of the negative log-likelihoods over all vocabulary positions:

$$\mathcal{L}_{ar} = -\frac{1}{T}\sum_{t=1}^{T} \log p(q_t|q_{<t}). \quad (8)$$

Through optimization of Eq. 8, the model studies the transformation process from the previous $t-1$ tokens to the t-th token within the current sequence. Due to the presence of position encoding, the model may be compelled to memorize the token transformation process on the basis of positional information, thereby potentially compromising its capacity for generalization.

Data Augmentation. In the aforementioned limitations where training data and generalization are limited, data augmentation emerges as an efficient approach to enhance the adequacy of model training. By harnessing the flexibility of perspectives inherent in the multi-view image generation task, we develop the Shuffle Views (ShufV) data augmentation strategy. Upon implementing ShufV, Eq. 3 can be reformulated as follows.

$$q^{(S_n-1)*h*w+(i-1)*w+j} = \left(\arg\min_{v\in[V]} \|\text{lookup}(Z, v) - f_n^{(i,j)}\|_2\right) \in [V], \quad (9)$$

where S denotes a random sequence. S_n indicates that the image of the n-th view is now utilized as the S_n-th view in the training sequence q.

ShufV has the potential to significantly increase the quantity of our training data manifold. In the context of training MV-AR with n views, ShufV can generate $\frac{n(n-1)}{2}$ permutations and combinations of these perspectives. Consequently, a single object or scene can be expanded into $\frac{n(n-1)}{2}$ training sequences, considerably enhancing the diversity and richness of the training data.

It is pertinent to observe that both self-attention and FFN exhibit the permutation equivariance. Consequently, alterations in the sequence order of inputs will induce corresponding changes in the sequence order of the model's intermediate features. To ensure that auxiliary conditions, such as camera poses and reference images, effectively guide the model in generating images in a predetermined sequence, it is imperative to rearrange these conditions. This

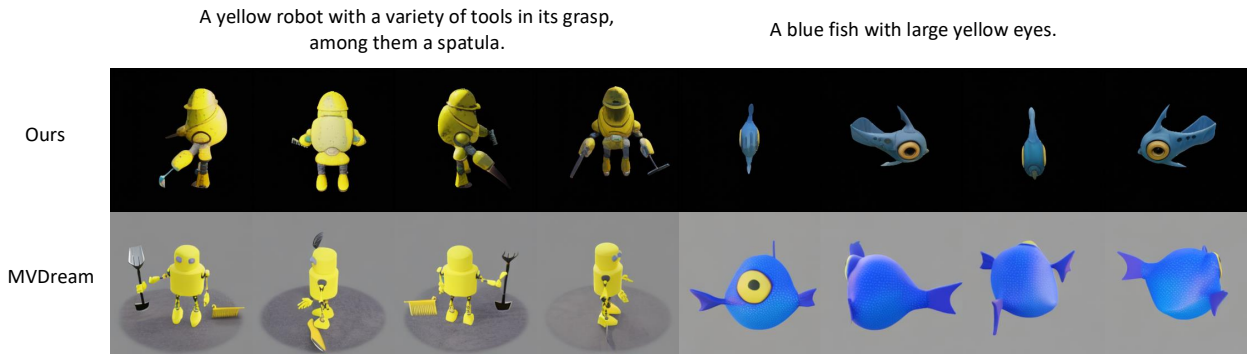

Figure 3. **Qualitative study of text to multi-view generation** with diffusion-based method [36].

reordering will ensure alignment of the condition sequence with that of the input sequence.

Discussion on ShufV. We argue that the ShufV also helps alleviate the part of **Issue 1**: the AR model has difficulty exploiting the overlapping conditions between successive and current views. Using ShufV for data augmentation, the order of views is not fixed. Considering that there are two views $\mathcal{A}$ and $\mathcal{B}$ in the input sequence q. ShufV enables the AR model to acquire the transformation from view $\mathcal{A}$ to view $\mathcal{B}$ during the training phase, as well as from view $\mathcal{B}$ to view $\mathcal{A}$ after using ShufV. It enhances the IWC's ability to capture the overlap between view $\mathcal{A}$ and view $\mathcal{B}$ with view $\mathcal{B}$ as a reference, and vice versa. Consequently, this permits the model to exploit the overlap conditions between the current view and other views and use them effectively.

Progressive Learning. Utilizing the aforementioned loss function Eq. 8 and the "ShufV" data augmentation Eq. 9, we can successfully train a text-to-multi-view (t2mv) model. This t2mv model employs SSA to analyze text input and efficiently produce multi-view images that correspond to the given text. We use the t2mv model as a baseline to train the X-to-multi-view (X2mv) model to facilitate the versatility of the MV-AR in alternative conditions, such as images and shapes.

During training of the X2mv model, the text condition is randomly dropped, while other conditions are randomly combined, as shown in Figure 2. When the text prompt is dropped, it is replaced by a statement that does not pertain to the target image. For example, a command such as "Generate multi-view images of the following <img>" may be utilized. In this context, "<img>" signifies that a reference image will be combined after the text. In cases where the subsequent element is a geometric shape, "<img>" is substituted with "<shape>". The probability of a condition dropping and combining increases linearly from 0 to 0.5 as the number of training iterations increases, and remains 0.5 in subsequent training. This escalation is confined to the initial 10k training iterations. This progressive learning allows the model to be influenced by new conditions introduced during training while maintaining a certain level of adherence to the text prompts.

4. Experience

To evaluate the performance and versatility of our MV-AR, we perform a comparative analysis in three specific tasks: (1) text-to-multi-view image generation (refer to Sec. 4.1), (2) image-to-multi-view image generation (refer to Sec. 4.2), and (3) shape-to-multi-view image generation (refer to Sec. 4.3). In each subsection, we establish ablation studies to evaluate the effectiveness of our designed architectures or training strategies.

4.1. Text to multi-view

Training details. Following MVDream [36], we train our MV-AR model on a subset of the objaverse [8] dataset, which comprises 100k high-quality objects. The text instructions are provided by Cap3D [25]. The training process involves 30k iterations on 16 A800 GPUs with batch-size 1,024 and learning rate 4×10^{-4}. The optimizer is AdamW with $\beta_1 = 0.9, \beta_2 = 0.95$.

Metrics. The Frechet Inception Distance (FID) [13] and the Inception Score (IS) [2] are used to assess image quality, while the CLIP-Score [12] is used to assess text-image consistency, with scores averaged across all views.

Results. We adopt the Google Scanned Objects [9] as our primary evaluation benchmark. To generate textual descriptions, we render each object from multiple viewpoints and leverage GPT-4V to annotate its visual attributes. A random selection of 30 objects is selected that included the daily items for the animals. For each selected object, an image of size 256×256 is rendered to serve as an input view.

Table 1 illustrates that our MV-AR exhibits a comparable multi-view image quality and text-image consistency consistency. Compared to the diffusion-based method [36], our MV-AR demonstrates equivalent image generation quality while possessing improved image-text consistency, as evidenced by the higher CLIP Score metric attributed to MV-AR. When evaluated against the baseline model LLamaGen [38], our approach markedly elevates the IS and

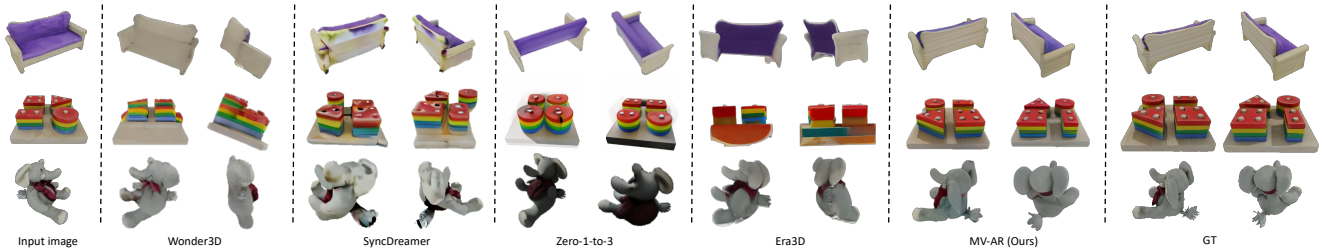

Figure 4. **Qualitative comparison on image-conditioned multi-view generation** with diffusion-based method [17, 21–23]. Era3D [17] generates results in canonical space, so there are certain view differences compared with other methods.

Methods	FID ↓	IS ↑	CLIP-Score ↑
MVDream [36]†	**141.05**	7.49	28.71
LLamaGen [38]†	146.11	5.78	28.36
MV-AR (Ours)	144.29	**8.00**	**29.49**

Table 1. **Ablation on text-condition module on GSO**. We report FID, IS, and CLIP-Score on the generated multi-view images of 30 GSO objects. † means the model is finetuned only on Objaverse subset for fair comparison.

CLIP-Score metrics due to refinement in processing the text condition from conventional self-attention to the proposed SSA. This demonstrates that our proposed SSA increases the congruence between image and text, as well as the overall quality of generation. The output images are also presented in Figure 3. Compared to MVDream [36], the images generated by our MV-AR exhibit a superior multi-view consistency, which is particularly evident in the consistency between the front-view and back-view images. The tool held in the right hand of the yellow robot as generated by MVDream displays significant deformation in front-view and back-view, whereas the yellow robot produced by MV-AR consistently maintains a similar object across all views.

4.2. Image to multi-view

Training details. Same as the text-to-multi-view (t2mv) task, the Objaverse subset is used for the training data. Multi-view images, which are arrayed around a circular configuration of objects, are selected as the training images. During training of the 4-view image generation model, the angular difference between each view is set to 90 degrees. Hyper-parameters in training process, such as iterations, learning rate and optimizer, are the same as the t2mv task.

Metrics. We employ widely used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Metric (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [46], for the image to multi-view generation task. These metrics are effective in quantitatively assessing the consistency of color and texture between the output image and the corresponding ground-truth image.

Ablation on image condition. The experimental result in Table 2 indicates that cross-attention mechanisms, although widely used within diffusion-based methods [22, 36], present significant challenges when applied to AR frameworks. This difficulty is attributed to the base model employed, which lacks inherent image-to-image capabilities. In contrast, the foundational model underpinning diffusion-based methods, *e.g.*, SD2 [33] and SDXL [30], encompasses specific image-to-image functionalities acquired through extensive pre-training. Unlike cross-attention mechanisms, our IWC enables accurate output regulation by incorporating condition token-by-token into the model, thereby facilitating effective control.

Image condition	PSNR ↑	SSIM ↑	LPIPS ↓
In-context	11.92	0.538	0.477
Cross Attention	15.13	0.709	0.310
IWC (Ours)	**22.99**	**0.907**	**0.084**

Table 2. **Ablation on image-condition module on GSO.** We report PSNR, SSIM, and LPIPS on the generated multi-view images of 30 GSO objects.

Evaluation generalization on GSO. We evaluate the generalization of our MV-AR in image-to-multi-view (i2mv) tasks in Table 3. We adopt RealFusion [26], Zero123 [21], SyncDreamer [22], Wonder3D [23], and Era3D [17] as comparison methods. Given an input image, these methods generate images of multiple specific views. The quantitative results demonstrate that our MV-AR attains the highest PSNR and the second highest SSIM. These results suggest that our method is more effective in preserving the color and texture fidelity of the reference image, leading to a more consistent generation of multiple views. The LPIPS of our MV-AR is positioned third among all methods, this result indicates that the MV-AR may be overly stringent in enforcing the consistency of low-level features between the generated and reference images, which consequently diminishes the perceptual quality [46].

A qualitative comparative analysis of our method against Wonder3D [23] and SyncDreamer [22] is presented in Figure 4. We illustrate the superiority of autoregressive multi-view image generation over diffusion-based simultaneous multi-view generation through a scenario where generating back-view images from the front view. MV-AR effectively

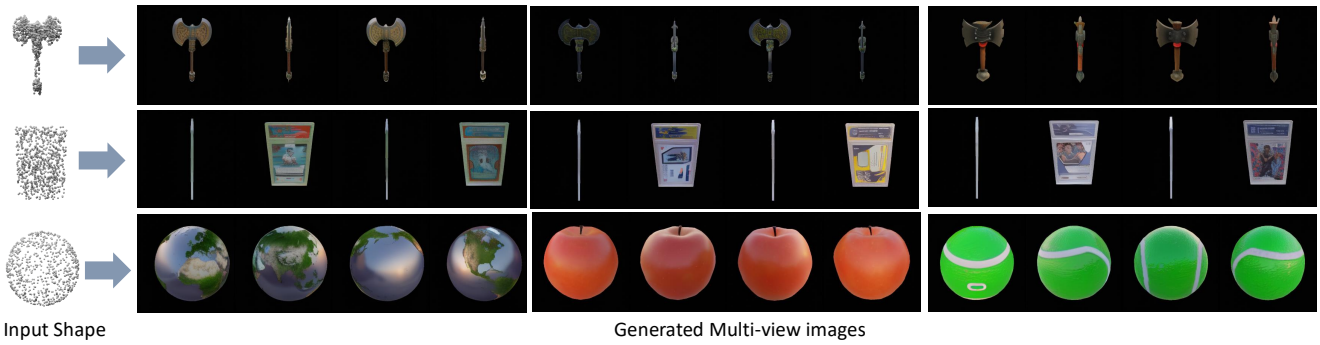

Figure 5. **Visualization of our shape-conditioned multi-view generation.** Given a geometric shape, our approach robustly generates multi-view images that are geometrically consistent with it, while ensuring multi-view coherence.

Image condition	PSNR ↑	SSIM ↑	LPIPS ↓
Realfusion [26]	15.26	0.722	0.283
Zero123 [21]	18.93	0.779	0.166
SyncDreamer [22]	19.89	0.801	0.129
Wonder3D [23]	22.82	0.892	**0.062**
Era3D [17]	22.73	**0.911**	0.071
MV-AR (Ours)	**22.99**	0.907	0.084

Table 3. **Quantitative results on image-conditioned multi-view synthesis on GSO**. We report PSNR, SSIM, and LPIPS on the generated multi-view images of 30 GSO objects.

utilizes the cumulative information from all preceding viewpoints, as opposed to relying on a singular view. This progressive generation allows for a comprehensive reference during generation, leading to better qualitative results.

4.3. Shape to multi-view

Training details. The training data and hyper-parameters used are consistent with those used in the image-to-multi-view task. Furthermore, we implement our proposed progressive learning approach to enhance the model's ability in multi-modal conditional processing.

Results. To validate the efficacy of our shape-conditioned multi-view generation framework, we fix the shape token and generate multi-view outputs across multiple trials, as illustrated in Figure 5. Our method consistently produces outputs that not only adhere to the input shape constraints with high precision but also exhibit semantically plausible and diverse textures.

4.4. Ablation study

	FID ↓ / IS ↑	PSNR ↑ / SSIM ↑ / LPIPS ↓
w/o SPE	147.29 / 7.26	21.30 / 0.843 / 0.118
w/o ShufV	173.51 / 4.77	18.27 / 0.778 / 0.194
MV-AR (Ours)	**144.29 / 8.00**	**22.99 / 0.907 / 0.084**

Table 4. **Ablation study of SPE and ShufV**, with the performance on the t2mv and i2mv tasks presented simultaneously.

Effect of SPE camera pose guidance. We demonstrate that using the camera pose as the shift position embedding is helpful in generating multi-view consistent images.

Effect of ShufV data augmentation. We show that the ShufV data augmentation strategy can significantly improve the quality of the output images.

5. Conclusion

In this study, we introduce an approach for auto-regressively generating multiview images. The motivation is to ensure that, during the generation of the current view, the model can extract efficient guidance information from all preceding views, thus enhancing the multi-view consistency. Based on this idea, we develop the MV-AR model, which progressively produces multi-view images based on multi-modal conditions. Specifically, we initially design several conditional injection modules for four conditions: text, image, camera pose, and geometry. Subsequently, we implement progressive learning to enable the MV-AR to manage these conditions concurrently. Finally, we employ "Shuffle View" data augmentation to expand the dataset, mitigating the inherent overfitting issues associated with the AR model when applied to limited training data. As a result of these novel designing, our MV-AR model achieves generation performance that is comparable to the state-of-the-art diffusion-based multi-view image generation and establishes itself as the first multi-view image generation model capable of handling multi-modal conditions.

Future work. We will enhance our MV-AR from: *(1) Better tokenizer*. The reason why we do not use 3D VAE is the exchange of information between views during its encoding, which contradicts the core motivation of our study. Therefore, we will focus on improving performance via tokenizing multi-view images using causal 3D VAE. *(2) Unified generation and understanding*. This study employs the AR model to accomplish the multi-view image generation task. In future work, we aim to harness the comprehensive capabilities of AR to unify the processes of multi-view generation and understanding.

Acknowledgement

This work was supported financially the Natural Science Foundation of China (82371112, 623B2001), the Science Foundation of Peking University Cancer Hospital (JC202505), Natural Science Foundation of Beijing Municipality (Z210008) and the Clinical Medicine Plus X - Young Scholars Project of Peking University, the Fundamental Research Funds for the Central Universities (PKU2025PKULCXQ008).

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. 3

[2] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 6

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[4] 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. 3

[5] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. Reversible column networks. In *The Eleventh International Conference on Learning Representations*, 2022. 5

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 4

[8] 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. 3, 6

[9] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 6

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[11] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 3

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 4, 6

[13] 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. *Advances in neural information processing systems*, 30, 2017. 6

[14] JiaKui Hu, Lujia Jin, Zhengjian Yao, and Yanye Lu. Universal image restoration pre-training via degradation classification. *The Thirteenth International Conference on Learning Representations*, 2025. 5

[15] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10026–10038, 2024. 2, 4, 5

[16] 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. 3

[17] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wei Xue, Wenhan Luo, et al. Era3d: high-resolution multiview diffusion using efficient row-wise attention. *Advances in Neural Information Processing Systems*, 37:55975–56000, 2024. 1, 7, 8

[18] Sixu Li, Chaojian Li, Wenbo Zhu, Boyang Yu, Yang Zhao, Cheng Wan, Haoran You, Huihong Shi, and Yingyan Lin. Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–13, 2023. 2

[19] Jialun Liu, Chenming Wu, Xinqi Liu, Xing Liu, Jinbo Wu, Haotian Peng, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Texoct: Generating textures of 3d models with octree-based diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4284–4293, 2024. 2

[20] Jialun Liu, Jinbo Wu, Xiaobo Gao, Jiakui Hu, Bojun Xiong, Xing Liu, Chen Zhao, Hongbin Pei, Haocheng Feng, Yingying Li, et al. Texgarment: Consistent garment uv texture generation via efficient 3d structure-guided diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26566–26575, 2025. 2

[21] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2, 7, 8

[22] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 7, 8

[23] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1, 2, 7, 8

[24] Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8744–8753, 2024. 2

[25] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36: 75307–75337, 2023. 6

[26] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8446–8455, 2023. 7, 8

[27] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International conference on machine learning*, pages 7220–7229. PMLR, 2020. 3

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 5

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023. 7

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[32] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2

[33] 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. 1, 7

[34] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 4

[35] 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, 2023. 1, 2

[36] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 2, 6, 7

[37] 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. 3

[38] 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. 3, 6, 7

[39] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025. 3

[40] 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. 4

[41] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 3

[42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[44] Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 551–561, 2025. 2

[45] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[46] 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. 7

[47] 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*, 36:73969–73982, 2023. 5

[48] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Liefeng Bo, Zilong Dong, Qixing Huang, et al. Videomv: Consistent multi-view generation based on large video generative model. 2024. 1