AR-1-to-3: Single Image to Consistent 3D Object via Next-View Prediction

Xuying Zhang[1*] Yupeng Zhou[1*] Kai Wang[1] Yikai Wang[3] Zhen Li[1]
Shaohui Jiao[4] Daquan Zhou[4] Qibin Hou[1,2†] Ming-Ming Cheng[1,2]
[1] VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian [3]Tsinghua University [4]ByteDance Inc.
Homepage: https://github.com/HVision-NKU/AR123

Figure 1. Examples of high-quality 3D asset gallery generated by our AR-1-to-3 model. See Sec. 4.2 for more visualization results.

Abstract

Novel view synthesis (NVS) is a cornerstone for image-to-3d creation. However, existing works still struggle to maintain consistency between the generated views and the input views, especially when there is a significant camera pose difference, leading to poor-quality 3D geometries and textures. We attribute this issue to their treatment of all target views with equal priority according to our empirical observation that the target views closer to the input views exhibit higher fidelity. With this inspiration, we propose AR-1-to-3, a novel next-view prediction paradigm based on diffusion models that first generates views close to the input views, which are then utilized as contextual information to progressively synthesize farther views. To encode the generated view subsequences as local and global conditions for the next-view prediction, we accordingly develop a stacked local feature encoding strategy (Stacked-LE) and an LSTM-based global feature encoding strategy (LSTM-GE). Extensive experiments demonstrate that our method significantly improves the consistency between the generated views and the input views, producing high-fidelity 3D assets.

*Equal contribution. †Corresponding author.

1. Introduction

Synthesizing 3D objects from a single-view image has been a fundamental and enduring research in both computer vision and graphics communities. This task is challenging, as it not only involves reconstructing the visible regions of a 3D object but also necessitates extrapolating the invisible parts. To address these challenges, researchers have sought to synthesize novel views [4, 29, 32, 40] with the powerful priors of diffusion models, which have achieved remarkable success in image generation [7, 25, 38, 43, 74].

Existing approaches can be broadly categorized into two distinct paradigms. To be specific, the methods represented by Zero123 [30] fine-tune pre-trained diffusion models on <*input_view, camera_pose, output_view*> triplets of 3D renders and generate multiple discrete views of a 3D object based on a single-view image and a set of camera poses. On another line, Zero123++ [45] and its follow-ups [21, 28, 50, 65] consider the independence of new views specified by different camera poses. They utilize diffusion models to generate multiple views simultaneously by merging all target views with specified camera poses into a grid image for joint distribution modeling. Despite their success in many scenarios, these methods remain prone to synthe-

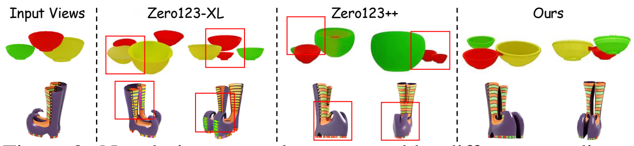

Figure 2. Novel view examples generated by different paradigms when encountering significant differences in camera poses.

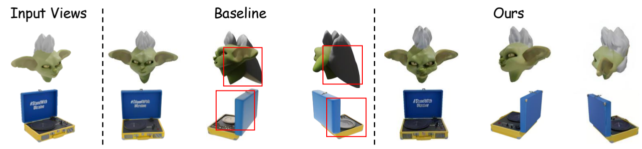

Figure 3. Novel view examples of nearby and distant camera poses generated by the Zero123++ baseline and our AR-1-to-3.

sizing new views inconsistent with the input image, especially when there is a large gap between their camera poses, as illustrated in the bowls and shoes of Fig. 2.

A fundamental reason for this issue is that these methods generate novel views for all camera poses in an equally prioritized manner. We argue that novel views with camera poses closer to the input view should be generated first in that they usually exhibit substantially higher generation fidelity, whereas those with larger pose variations present greater challenges, as shown in Fig. 3. To this end, we rethink the way to generate consistent views and present AR-1-to-3, a novel paradigm to progressively generate all target views, with the closer views generated first serving as contextual information to generate the farther view.

Our methodology follows the 3×2 grid image generation strategy proposed by Zero123++ [45]. It is noteworthy that there exists a potential sequential relationship between these six target views, where the adjacent rows of views in the grid share identical elevation angles and a fixed azimuth interval of $120°$. This sequential nature enables our AR-1-to-3 to start from generating the first row views based on the input view and gradually generate the remaining row views in an autoregressive fashion. At each iteration step, the target views from two different camera elevations can exchange information, and the views generated in the previous steps are utilized as references to generate current views.

To encode partially generated sequence views and provide references for the next views, we develop two image conditioning strategies, *i.e.,* Stacked-LE and LSTM-GE, for local conditioning and global conditioning of the diffusion models respectively. In the Stacked-LE strategy, the denoising UNet model encodes the previously generated views into a stack embedding, which is employed as pixel-wise spatial guidance to modify the key and value matrices of the self-attention layers for denoising the target views of the current step. In LSTM-GE, the view subsequence is divided into two groups according to the elevation, whose feature vectors are encoded by two LSTM modules for high-level semantic conditions of current views.

We evaluate the performance of our AR-1-to-3 paradigm on a 3D benchmark dataset, *e.g.,* Objaverse [9], and two out-of-domain datasets, *i.e.,* Google Scanned Objects [10], and OmniObject3D [63]. By introducing the autoregressive manner coupled with the proposed Stacked-LE and LSTM-GE strategies into multi-view diffusion models, our AR-1-to-3 enables consistent and accurate novel view synthesis, resulting in high-quality 3D assets, as shown in Fig. 1. The experimental results also demonstrate the superiority of our AR-1-to-3 compared to cutting-edge new view synthesis methods and image-to-3d methods.

Our contributions can be summarized as follows:

- We propose AR-1-to-3, an autoregressive next-view prediction framework for 3D object generation, which can generate target views from near to far progressively.
- We design Stacked-LE and LSTM-GE strategies to encode the partially generated sequence views and provide local and global conditions for the diffusion models.
- Extensive quantitative and qualitative experiments on large-scale 3D datasets demonstrate that our approach can generate more consistent 2D multi-view images than previous works and produce high-quality 3D assets.

2. Related Work

2.1. 2D Diffusion Models for 3D Generation

Diffusion Models [8, 11, 20, 38, 43] pre-trained on large-scale 2D datasets have demonstrated remarkable performance in generating high-quality images and powerful zero-shot generalization. In recent years, significant effort has been consecutively devoted to transferring such merits of 2D diffusion models to 3D generation.

Early methods [2, 23, 39, 41, 71] distill knowledge of 2D pre-trained models by feeding the rendered views to it and performing per-shape optimization but suffer from artifacts such as over-saturated colors and the “multi-face” problem. Zero123 [30] pioneers an open-world single-image-to-3D framework, in which diffusion models are fine-tuned to synthesize new views conditioned on an input view and a set of discrete camera poses. ImageDream [60] adopts world camera coordination as in MVDream [47] to recover 3D geometries. Magic123 [40] combines the 3D prior of Zero123 and the 2D prior of the stable diffusion model together to enhance the quality of generated 3D meshes. One-2-3-45 [29] uses Zero123 to generate multi-view images, which are lifted to 3D space to assist in the generation of 3D meshes. Several approaches like Consistent123 [27] and MVD-Fusion [17], improve Zero123 by incorporating additional priors, *i.e.,* boundary and depth. Besides, more and more attention has been drawn to enforce consistency between the generated multiple views. SyncDreamer [31] adopts a 3D-aware attention mechanism to correlate the corresponding features across different views. MVDiffusion [51] generates multi-view images in parallel through the weight-sharing multi-branch UNet with shared weights and correspondence-aware attention. Recent methods [13, 22, 52, 73] explore to simultaneously generate multiple views to model their joint distribution. Zero123++ [45] further proposes a strategy of tiling six target views surrounding the 3D object into a grid image. This strategy has been successively adopted by follow-up works such as One-2-3-45++ [28], Instant3D [21], and InstantMesh [65].

In contrast to these methods heavily relying on the 2D diffusion priors, our AR-1-to-3 inspired by human thinking also pays attention to the contextual information of the current object. The method most similar to ours is Cascade-Zero123 [4]. It first uses a multi-view diffusion model to generate many extra views, which, along with the input image, are then fed into another diffusion model to produce a specific target view. Different from this method, our AR-1-to-3 takes the relationship between the target views and the input image into account and utilizes a diffusion model to establish the potential sequence between them.

2.2. Autoregressive Generation

The autoregressive scheme, which operates on the fundamental principle that the current value in a sequence can be explained by its preceding values, plays a vital role in the deep learning era [35, 37, 59, 64, 69]. Based upon this technology, researchers have developed a series of classical sequential modeling methods [15, 26, 36, 46, 57]. In recent years, there has been a growing interest in extending such an autoregressive fashion to various communities.

PixelRNN [55] is one of the pioneering methods to generate high-quality images with intricate details by modeling pixel dependencies. VQ-VAE [56] revolutionizes the learning of discrete representations by incorporating the codebook mechanism, enabling efficient encoding and decoding processes of images. VQ-GAN [12] adopts a transformer architecture to model serialized visual parts and introduces adversarial loss during the training process. Parti [66] proposes a pathways autoregressive model treating the image generation as a sequence-to-sequence modeling to generate high-fidelity photorealistic images. VAR [53], LlamaGen [49], and Infinity [14] scale image generation by incorporating multimodal large models. In addition, some approaches [1, 19, 44, 58, 77] attempt to integrate diffusion models with the sequential strategy to achieve more temporally consistent video generation. More recently, MeshGPT series [3, 5, 48] propose to generate triangle faces in an autoregressive manner for artist-created 3D meshes. TAR3D [70] and SAR3D [6] quantify 3D latent representations [18], which are used to generate SDF-based 3D objects [75, 76] with the next-token prediction strategy.

In this work, we observe the target views of Zero123++ can be divided into several steps with equidistant camera pose intervals, resulting in a potential sequential nature.

3. Methodology

3.1. Preliminaries

We introduce the preliminaries of Zero123++ [45], the base multi-view diffusion model adopted in this work, which is beneficial for understanding the designs in AR-1-to-3.

Multi-View Generation. To achieve the modeling of the joint distribution between multiple new views, Zero123++ proposes to tile six target views with a 3×2 layout in a grid image. Note that these target views are obtained from a fixed set of relative azimuth and absolute elevation angles. Specifically, they consist of interleaving elevation angles of $20°$ downwards and $-10°$ upwards, combined with azimuth angles starting from $30°$ relative to the input azimuth and incrementing by $60°$ for each subsequent camera pose.

Stable Diffusion. Zero123++ chose Stable Diffusion (SD) as the generative model since it is open-sourced and has been trained on various internet-scale image datasets. The geometric priors that SD learns about natural images are utilized for novel view synthesis under the image and camera conditions. SD performs the diffusion process within the latent space of a pre-trained autoencoder whose encoder and decoder are denoted as $\mathcal{E}(\cdot)$and $\mathcal{D}(\cdot)$, respectively. At the diffusion time step t, the objective for fine-tuning the denoiser UNet $\epsilon_\theta(\cdot)$ can be formulated as:

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),y,t,\epsilon\sim\mathcal{N}(0,I)}||\epsilon - \epsilon_\theta(z_t, t, c_\theta(y)||_2^2, \quad (1)$$

where x is the target grid image which is perturbed to a feature by Gaussian noise ϵ, *i.e.,* z_t, and $c_\theta(y)$ represents the embedding encoded from the condition image y.

Image Condition. The image conditioning techniques in Zero123++, *i.e.,* $c_\theta(y)$, can be divided into local and global conditions. The former strategy is tailored for the pixel-wise spatial correspondence between the input and target views. Specifically, it adopts a variant of the Reference Attention operation [67], which runs the denoising UNet on the input image and appends the self-attention key and value matrices from it to the corresponding attention layers when denoising the target views. Note that the Gaussian noise at the same level as the denoising input is added to the referring image so that the UNet can focus on the relevant features for denoising at the current noise level. In terms of the global conditioning strategy, the CLIP [42] text embedding of an empty text is added with the CLIP image embedding of the input image multiplied by a trainable set of global weights to provide high-level semantic information for the cross-attention of the denoising UNet.

3.2. AR-1-to-3

Existing methods either generate multiple discrete viewpoints from a single input view and a set of camera poses, like Zero123 [30], One-2-3-45[29], or simultaneously generate multiple views in a grid layout based on specified cam-

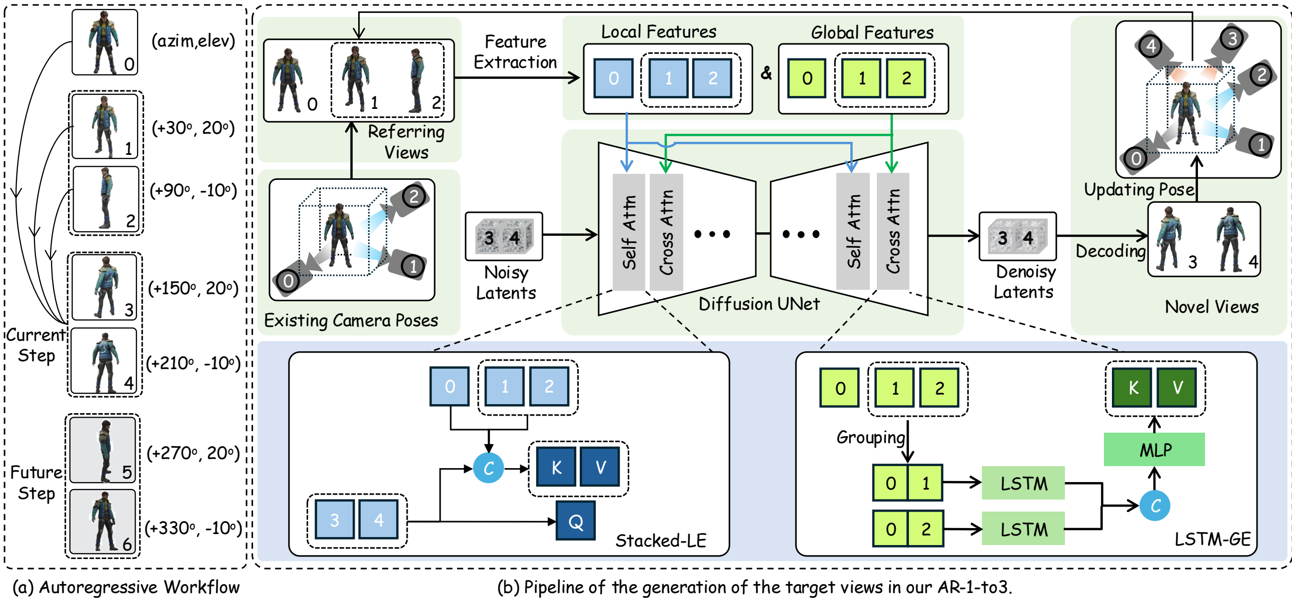

Figure 4. Overview of our AR-1-to-3 framework. The left side shows the AR-1-to-3 workflow, while the right side illustrates the denoising process of target views. Taking the input single-view image as initialization, our methodology employs a diffusion model to generate all target views incrementally from near to far, with the existing views from previous steps serving as contextual information about the objects themselves. To achieve this, the Stacked-LE and LSTM-GE strategies are developed to encode the local features and global features of the partial view sequence as the image conditions of the denoising UNet for the view prediction of the current step.

era conditions, such as Zero123++ [45], ImageDream [60]. Despite achieving excellent performance in many scenarios, these methods are still prone to generating several target views that exhibit geometric and textural inconsistencies with the input image. We argue that their equal prioritization of all new views and the underutilization of contextual information of current objects during the generation process should be responsible for this issue.

The core of this work is to progressively generate target views in a next-view prediction fashion so that the closer views generated earlier can be used as supplementary information for the generation of the farther views. Fig. 4 shows the end-to-end architecture of our AR-1-to-3. In this work, we follow the paradigm of Zero123++ generating 6 specific camera conditional target views. In contrast, our method generates these views step by step rather than all at once. In addition, Zero123++ has demonstrated that generating multiple target views simultaneously contributes to accurately modeling the joint distribution of these views. Therefore, each step in our generative strategy refers to a row of the 3×2 layout, containing two target views with different elevations and 60° difference in azimuth. Moreover, the difference in camera pose between adjacent steps is a fixed azimuth angle of 120°, which is suitable for the next view prediction scheme. As a result, the different target views at each step can exchange information, and the target views generated in the previous steps can be utilized as extra conditions to generate the views for the current step.

We achieve such next-view prediction by designing two image conditional strategies that encode sequence view information to fine-tune the diffusion model. These two strategies, denoted as Stacked Local Feature Encoding (Stacked-LE) and Long Short-Term Global Feature Encoding (LSTM-GE), correspond to the local and global image conditioning techniques in Zero123++, respectively. The optimization objective can still be represented by Eqn. 1, and we will elaborate on our image conditional policy, *i.e.*, $c_\theta(y)$, in the subsequent Sec. 3.3 and Sec. 3.4.

Through multi-step autoregression, our AR-1-to-3 gradually generates 6 target views, which are fed to a sparse-view large reconstruction model to obtain a 3D object. In this work, we choose the pre-trained InstantMesh [65] as our 3D reconstruction model, which encodes the multi-view images as triplane features via a transformer-based architecture [16] and predicts the point color and density for volumetric rendering by a multi-layer perceptron.

3.3. Stacked Local Feature Encoding

In this subsection, we introduce how our method encodes the latent features of the input image and the generated partial target-view sequence as local conditions for the operation of Reference Attention to generate the target views of the current step. Note that the denoising UNet model is a multi-level architecture and the hidden dimensions may vary across different self-attention layers. It is challenging to encode the latent features of the condition view sequence at these positions using a single network. Considering that these reference features and the attention representations in self-attention layers come from the same positions in the U-Net network, sharing the same spatial and channel dimensions, we naturally thought of encoding them into a unified representation by stacking them along the spatial dimension. This strategy offers two significant advantages: 1) it can encode any number of reference features at the self-attention layer. 2) it allows the reference features to be directly fed into the attention module, enabling the reuse of weight parameters without any additional design required. We term this local feature encoding strategy as Stacked-LE whose details are shown in the bottom left corner of Fig. 4.

Formally, at the k-th step of autoregression, with a total of $2k$-1 reference views, we aim to predict the $(2k)$-th and $(2k$+1$)$-th target views, where the variable k ranges from 1 to 3. Following Zero123++ [45], we first feed the referring views into the denoising UNet model separately and record their key/value matrices at the self-attention module. Then, we perform another forward pass with the U-Net to denoise the target views of the current step. During this process, the records in each layer are stacked together to modify the key and value matrices in the self-attention module of the corresponding layer, which can be defined as:

$$s_i^* = \text{Concat}([e_i^1, e_i^2, ..., e_i^{2k-1}, s_i]), \quad (2)$$

where $s_i \in \mathbb{R}^{B \times L \times D_i}$ denotes the key or value matrices of the i-th self-attention, and e_i^j is the recorded embeddings of the j-th reference view. Note that B, L and D_i are the batch size, token number, and dimension of the key /value matrices. Further, we compute the self-attention as follows:

$$O_i = \text{Attention}([Q_i, K_i^*, V_i^*]). \quad (3)$$

3.4. Long Short-Term Global Feature Encoding

In this subsection, we present the details of encoding the CLIP features of the input images and the existing target-view sequence as global conditions, which provide high-level semantic information via cross-attention to generate the target views of the current step. We empirically observe that the CLIP features of the conditional view sequence are vectors with the same channel dimension and 1D spatial dimensions. In addition, they can be divided into two sub-sequences with an azimuthal angle spacing of 120° based on their elevation angles. Thus, it is well-suited to process such sequences with the Long Short-Term Memory (LSTM) Network [15]. Furthermore, this manner has two interesting merits, mitigating computational burden: 1) It can encode feature vectors from the conditional views into two vectors, regardless of the number of views. 2) The LSTM structure possesses a strong ability to model sequential representations while requiring fewer parameters. This strategy is named as LSTM-GE, and the diagrammatic details are shown in the bottom left corner of Fig. 4.

Given $2k$-1 conditional views at the k-th step of autoregression, we first send them to the image encoder of the CLIP model for their visual features, which are represented as $F \in \mathbb{R}^{B \times (2k-1) \times D}$. Then, we partition these features into two groups according to the elevation angles of their respective views. Note that the feature of the input image is a special case, which is included in both groups. We denote the two grouped features as $F_1 \in \mathbb{R}^{B \times k \times D}$ and $F_2 \in \mathbb{R}^{B \times k \times D}$, and feed them to two separate LSTM modules. Next, the hidden states of the k-th step of the two LSTM modules, *i.e.,* h_l^k, are selected as their respective outputs, *i.e.,* , I_l. This process can be formulated as follows:

$$I_l = \text{LSTM}_l([F_l, (h_l^0, c_l^0)]), \quad l \in \{0, 1\} \quad (4)$$

where h_l^0 and c_l^0 are the hidden state and cell state, respectively, both initialized as zero vectors. Finally, the outputs of the two LSTM modules are concatenated together along the channel dimension, followed by an MLP layer and a trainable set of global weights $W \in \mathbb{R}^{77 \times 1}$, to form global embeddings for the cross-attention of the denoising UNet:

$$T = W \cdot \text{MLP}(\text{Concat}([I_0, I_1])). \quad T \in \mathbb{R}^{B \times 77 \times D} \quad (5)$$

Note that we remove the CLIP embedding of empty text from the original global condition, as empirical observations indicate its negligible impact on the final results.

4. Experiments

4.1. Experiment Setup

Datasets. We conduct experiments on the most popular benchmarking 3D dataset, *i.e.,* Objaverse [9], along with two out-of-domain datasets, *i.e.,* Google Scanned Objects (GSO) [10] and OmniObject3D (Omni3D) [63]. Following the filtering principle in previous works [24, 70], we obtain about 210,000 geometry objects from the Objaverse dataset with a 3D mesh capacity of 800,000. We pick out 300 objects covering various categories to evaluate the performance of the methods, and the remaining samples are used for training. Moreover, we also randomly select 300 samples each from both GSO and Omni3D to ensure a fair evaluation of the methods involved in this paper.

Following the protocol of Zero123++ [45], we render 7 images for each object, including an input image and 6 target images. To be specific, an elevation angle ranging from $-20°$ to $45°$ and an azimuth angle ranging from $0°$ to $360°$ are randomly sampled to render the input image. The camera poses of the 6 target images involve interleaving absolute elevations of $20°$ and $-10°$, paired with azimuths relative to the input image that start at $30°$ and increase by $60°$ for each pose. Besides, all rendered images are set to a white background to ensure that the diffusion model produces images of this nature, thereby avoiding the trouble of

Figure 5. Visual comparisons of the novel views synthesized by our AR-1-to-3 and recent popular methods on multi-view generation. Compared with the existing approaches, the new views from our AR-1-to-3 are more consistent with both each other and the input views.

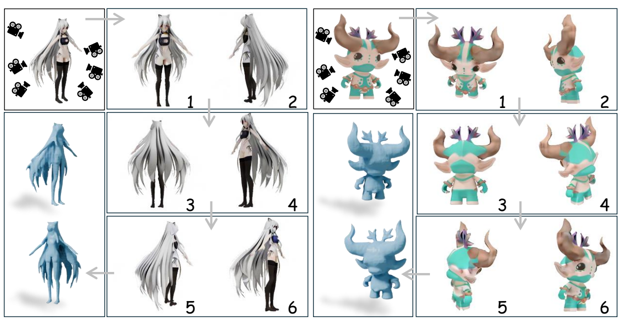

Figure 6. Examples of image-to-3d generation based on our next-view prediction. AR-1-to-3 produces multi-view images consistent with the input images, leading to high-quality 3D results.

removing the background when reconstructing 3D objects. We will open-source these rendered images in our project.

Evaluation. We evaluate the performance of the methods in two critical dimensions, *i.e.,* 2D visual fidelity and 3D geometric accuracy. Specifically, we compare the novel views generated by multi-view diffusion models or rendered from synthesized 3D meshes with the ground truth views. Following image comparison [4, 27, 65], we report four popular metrics, including Peak Signal-to-Noise Ratio (PSNR), Perceptual Loss(LPIPS) [68], Structural Similarity (SSIM) [61], and CLIP-score [42]. We also compare the surface points randomly sampled from the generated 3D meshes with those from ground-truth meshes, in which the chamfer distance (CD) value and the F-Score with a threshold of 0.02 are employed as metrics.

Implementation Details. We train AR-1-to-3 on the render images from about 210K objects of the Objaverse dataset for 150k steps with a total batch size of 32 on 8 NVIDIA A100 (80G) GPUs. The learning rate is initialized as $1e$-5 and changes every 25k steps in a cycle, along with the AdamW optimizer [34] and CosineAnnealingWarmRestarts scheduler [33]. We randomly select a k from $\{1, 2, 3\}$ to build the autoregressive pattern, where the first $2k$-1 views form the conditional images and the following two views make up the target views. We resize the size of the conditional images to a value between 128 and 512 so that the model is capable of adjusting to different input resolutions and producing more clear images. Meanwhile, we resize each target view to 320, thereby the size of the grid image during the autoregressive process is 320×640. In addition, we employ the linear noise schedule and v-prediction loss in Zero123++ [45] rather than the alternatives in the Stable Diffusion model [43]. During the inference stage, taking the input image as initialization, our AR-1-to-3 generates all target views in three steps, as shown in Fig. 6.

4.2. Qualitative Results

We perform qualitative experiments on novel view synthesis and image-to-3d on a wide range of 3D objects. To highlight the advantages of our AR-1-to-3 method in contextual reasoning and zero-shot generalization, we adopt input views with a certain degree of offset relative to the frontal views of 3D samples. We also provide more visualization results in the supplementary material.

Novel View Synthesis. Fig. 5 shows the synthesized views of our AR-1-to-3 and recent popular methods in multi-view generation, including Zero123-XL [30], SyncDreamer [31], Zero123++ [45], One-2-3-45 [29]. Note that Zero123-XL is an enhanced Zero123 model pre-trained on the Objaverse-XL dataset, and the open-source project of One-2-3-45 also employs this Zero123 version to generate the 8 views for its first stage. We utilize the elevation estimation implements of One-2-3-45 for the necessary elevation estimation procedures in Zero123-XL and SyncDreamer. Among

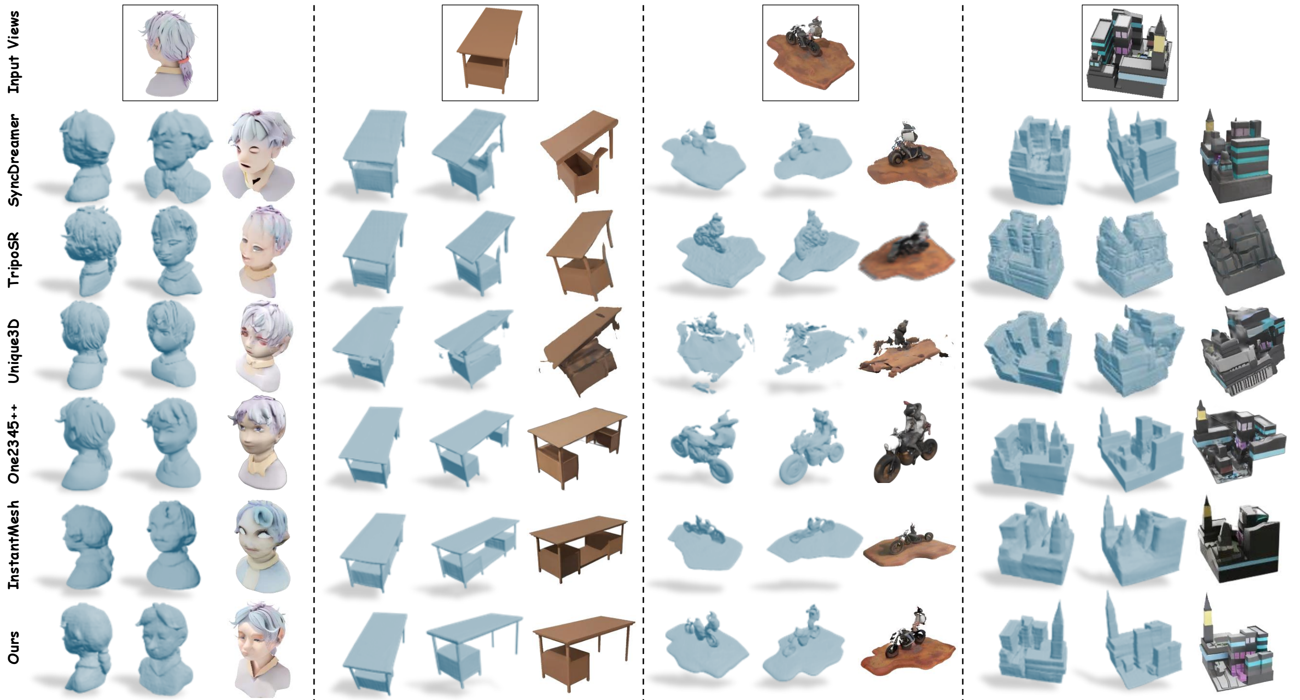

Figure 7. Visual comparisons between our AR-1-to-3 and recent cutting-edge methods on single-view image to 3D object generation. Note that the 3D results of Unique3D, TripoSR, and One2345++ are obtained by sending the input views to their official demos on Huggingface.

these difficult-to-maintain consistency scenes, some methods produce multiple inconsistent novel views, as shown in the bench results of Zero123++ and the cartoon figure of Zero123-XL. Some approaches may even be confused and generate new views that differ significantly from the input image, as shown in the bench predictions of SyncDreamer and the four-wheeled beds of One-2-3-45. In contrast, our AR-1-to-3 is able to capture texture details of 3D objects and synthesize consistent multi-view images, which can be attributed to the full utilization of contextual information.

Image-to-3D. Based on the synthesis of more consistent multiple views, our AR-1-to-3 can further generates high-quality 3D objects, as shown in Fig. 6. We also compare it with five cutting edge image-to-3d approaches, *i.e.,* SyncDreamer [31], InstantMesh [65], One2345++ [28], TripoSR [54], and Unique3D [62]. Note that the visual comparisons contain the pure geometries (left) and textured renderings (right) for each mesh generated by these methods. As depicted in Fig. 7, our AR-1-to-3 can generate 3D meshes with a consistent appearance and plausible geometry under limited input view information. Nevertheless, it is a real struggle for the counterparts to achieve this. For example, InstantMesh and One2345++ tend to generate an additional storage cabinet for the desk. We speculate that this is due to their excessive reliance on the symmetry prior of diffusion models during the generative process of new views, with less consideration for the contextual information of the object itself. Although SyncDreamer does not generate additional components, the desk it produces exhibits significant geometric deformations. We find that the reason stems from the inconsistency among the 16 views generated by its diffusion model, which is employed to reconstruct the 3D object. Unlike these approaches, our AR-1-to-3 effectively utilizes the contextual information of the object itself from nearby to distant during the autoregressive generation of all target views. As a result, our method can achieve excellent performance in image-to-3D generation.

4.3. Quantitative Results

We conduct quantitative studies on the two out-of-domain datasets, *i.e.,* GSO and Omni3D, to evaluate the performance of our AR-1-to-3 and other state-of-the-art methods fairly. Specifically, all candidate approaches are fed the same input images to generate 3D assets. We render 20 views with 224$\times$224 resolution for each mesh to perform the 2D evaluation. As far as the 3D evaluation, we uniformly sample 16K points from the mesh surfaces in an aligned cube coordinate system of $[-1,1]^3$. As shown in Tab. 1, our AR-1-to-3 surpasses other cutting-edge methods across all metrics. These results further demonstrate the superiority of our AR-1-to-3 in 3D asset creation.

4.4. Ablative Study

We conduct ablative experiments on the 300 objects excluded from training to examine the effectiveness of our key designs in the proposed AR-1-to-3 framework.

Table 1. Quantitative comparisons of our AR-1-to-3 model with the cutting-edge image-to-3d methods, including three 2D visual quality metrics and 3D geometric quality metrics. '↑': the higher the value, the better the performance, '↓': the lower the better.

Methods	GSO Dataset					Omni3D Dataset				
	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	F-Score ↑	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	F-Score ↑
Michelangelo [72]	9.323	0.609	0.408	0.165	0.105	9.969	0.602	0.410	0.174	0.081
SyncDreamer [31]	10.82	0.652	0.332	0.108	0.125	9.485	0.585	0.436	0.196	0.067
LGM [50]	9.139	0.592	0.429	0.157	0.075	10.02	0.588	0.394	0.152	0.086
InstantMesh [65]	10.67	0.661	0.338	0.117	0.135	9.91	0.608	0.412	0.178	0.076
AR-1-to-3 (Ours)	**13.18**	**0.709**	**0.232**	**0.063**	**0.258**	**10.25**	**0.629**	**0.388**	**0.148**	**0.097**

Table 2. Ablative study on conditional components of AR-1-to-3.

Stacked-LE	LSTM-GE	PSNR ↑	LPIPS ↓	SSIM ↑
		14.83	0.201	0.815
✔		17.59	0.174	0.833
	✔	17.91	0.170	0.836
✔	✔	20.28	0.121	0.857

Table 3. Ablative study on sequence orders during the autoregressive generation of AR-1-to-3.

Model	PSNR ↑	LPIPS ↓	SSIM ↑	CLIP-Score ↑
Reverse	20.19	0.124	0.851	0.882
Random	17.36	0.167	0.839	0.774
Normal (Ours)	20.28	0.121	0.857	0.887

Ablation on Image Conditioning Components. Starting from the Zero123++ baseline, we first incorporate Stacked-LE strategy, then LSTM-GE independently, and finally integrate both strategies. As shown in Tab. 2, they individually contribute to performance improvements, and their combined integration produces a greater improvement compared to the baseline method. These results demonstrate both the effectiveness of the two image conditioning strategies and the significant contribution of our next-view prediction in advancing the accuracy of novel view synthesis.

Ablation on Different Sequence Orders. We define the sequence order from near to far in terms of camera poses relative to the input viewpoint as the normal order. We also provide two variants of the sequence order, *i.e.,* reverse order and random order, to generate target views. To be specific, the reverse order refers to the view sequence whose cameras move from far to near relative to the input view. Meanwhile, the random order places the middle row of the 3×2 grid layout as the first position, followed by the remaining two rows. As shown in Tab. 3, our normal order achieves the best performance, while the random sequence performs the worst, demonstrating the effectiveness of modeling the target views in a sequential manner. Note that the reverse order achieves a similar performance to our normal order. We believe the reason for this is that the reversed sequence under Zero123++ setting can be seen as the camera moving from near to far in another circle direction, as the camera moves in a circular motion around the 3D object.

Encoding Strategy of Global Feature Sequence. To highlight the effectiveness of our LSTM-GE for global feature encoding of view sequence, we design a matrix multiplication ('matmul') variant to encode these features. Specifically, we stack these features into a matrix with shape $\mathbb{R}^{(2k-1)\times D}$. Meanwhile, we repeat the trainable weights in global condition $(2k\text{-}1)$ times to obtain a matrix with shape $\mathbb{R}^{77\times(2k-1)}$. We multiply these two matrices to obtain a new matrix with shape $\mathbb{R}^{77\times D}$, which is utilized as the key and value matrix for the cross-attention mechanism of the denoising UNet. As depicted in Fig. 8, with the incorporation of extra contextual information, the 'matmul' variant can generate more consistent and high-quality multi-view images compared to the baseline method, *i.e.,* Zero123++. Nevertheless, this variant may lead to bias in the global semantic understanding of 3D objects, such as the shapes of the book sample. In contrast, our strategy can generate multi-view images that are faithful to the shape and texture of objects in the input views. These experiments indicate the LSTM proposal can more effectively capture the high-level semantic information of the 3D objects.

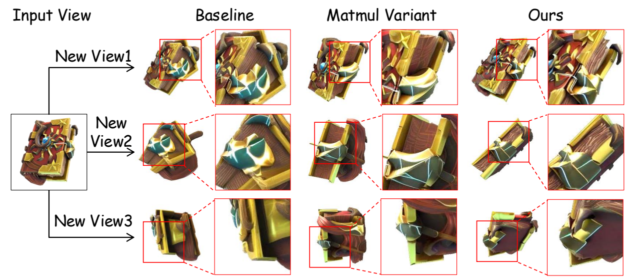

Figure 8. Ablative studies on the global feature encoding strategy.

5. Conclusions

In this paper, we present AR-1-to-3, a next-view prediction paradigm that starts from the input image and progressively generates target views from near to far. At each step of the autoregressive process, the previously generated views are employed as contextual information to facilitate the generation of the current target views. The experimental results demonstrate that our method generates new view images and 3D objects that are more consistent with the input images compared with the existing approaches generating multi-view images discretely or simultaneously.

Acknowledgments. This research was partially funded by Shenzhen Science and Technology Program (No. JCYJ20240813114237048). Besides, we express our gratitude to Dr. Xin Chen for his assistance in revising the paper.

References

[1] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2025. 3

[2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 2

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

[4] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. *arXiv preprint arXiv:2312.04424*, 2023. 1, 3, 6

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

[6] Yongwei Chen, Yushi Lan, Shangchen Zhou, Tengfei Wang, and XIngang Pan. Sar3d: Autoregressive 3d object generation and understanding via multi-scale 3d vqvae. *arXiv preprint arXiv:2411.16856*, 2024. 3

[7] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *arXiv preprint arXiv:2411.06558*, 2024. 1

[8] Zhennan Chen, Xuying Zhang, Tian-Zhu Xiang, and Ying Tai. Adaptive guidance learning for camouflaged object detection. *arXiv preprint arXiv:2405.02824*, 2024. 2

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 2, 5

[11] Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025. 2

[12] 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

[13] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 3

[14] 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

[15] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997. 3, 5

[16] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 4

[17] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 2

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

[19] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 3

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

[21] 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. 1, 3

[22] Xinyang Li, Zhangyu Lai, Linning Xu, Jianfei Guo, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Dual3d: Efficient and consistent text-to-3d generation with dual-mode multi-view latent diffusion. *arXiv preprint arXiv:2405.09874*, 2024. 3

[23] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. *Advances in neural information processing systems*, 37:75125–75151, 2024. 2

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

[25] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 1

[26] Zheng Li, Yibing Song, Penghai Zhao, Ming-Ming Cheng, Xiang Li, and Jian Yang. Atprompt: Textual prompt learning with embedded attributes. *arXiv preprint arXiv:2412.09442*, 2024. 3

[27] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023. 2, 6

[28] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024. 1, 3, 7

[29] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 6

[30] 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, 3, 6

[31] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 6, 7, 8

[32] 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

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

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

[35] Yuanxun Lu, Xinya Ji, Hao Zhu, and Xun Cao. Learning hierarchical adaptive code clouds for neural 3d shape representation. *Machine Intelligence Research*, 22(2):304–323, 2025. 3

[36] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023. 3

[37] Tai-Jiang Mu, Hao-Xiang Chen, Jun-Xiong Cai, and Ning Guo. Neural 3d reconstruction from sparse views using geometric priors. *Computational Visual Media*, 9(4):687–697, 2023. 3

[38] 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. 1, 2

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

[40] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1, 2

[41] Yansong Qu, Dian Chen, Xinyang Li, Xiaofan Li, Shengchuan Zhang, Liujuan Cao, and Rongrong Ji. Drag your gaussian: Effective drag-based editing with score distillation for 3d gaussian splatting. *arXiv preprint arXiv:2501.18672*, 2025. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 2021. 3, 6

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 6

[44] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024. 3

[45] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1, 2, 3, 4, 5, 6

[46] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 3

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

[48] 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

[49] 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

[50] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1, 8

[51] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023. 2

[52] Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Wangbo Yu, Chaoran Feng, Yatian Pang, Bin Lin, and Li Yuan. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. *arXiv preprint arXiv:2407.19548*, 2024. 3

[53] 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. 3

[54] 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. 7

[55] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 3

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

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

[58] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 3

[59] Haowei Wang, Jiayi Ji, Tianyu Guo, Yilong Yang, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Nice: improving panoptic narrative detection and segmentation with cascading collaborative learning. *arXiv preprint arXiv:2310.10975*, 2023. 3

[60] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2, 4

[61] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[62] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 7

[63] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2, 5

[64] Yu-Huan Wu, Da Zhang, Yun Liu, Le Zhang, and Ming-Ming Cheng. Ret3d: rethinking object relations for efficient 3d object detection. *Sci Sin Inform*, 55(4):887–901, 2025. 3

[65] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1, 3, 4, 6, 7, 8

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

[67] Lyumin Zhang. Reference-only control. *https://github.com/Mikubill/sd-webui-controlnet/ discussions/1236*, 2023. 3

[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[69] Tianyu Zhang, Guocheng Qian, Jin Xie, and Yang Jian. Fastpci: Motion-structure guided fast point cloud frame interpolation. In *ECCV*, 2024. 3

[70] Xuying Zhang, Yutong Liu, Yangguang Li, Renrui Zhang, Yufei Liu, Kai Wang, Wanli Ouyang, Zhiwei Xiong, Peng Gao, Qibin Hou, et al. Tar3d: Creating high-quality 3d assets via next-part prediction. *arXiv preprint arXiv:2412.16919*, 2024. 3, 5

[71] Xuying Zhang, Bo-Wen Yin, Yuming Chen, Zheng Lin, Yunheng Li, Qibin Hou, and Ming-Ming Cheng. Temo: Towards text-driven 3d stylization for multi-object meshes. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 19531–19540, 2024. 2

[72] 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. 8

[73] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9720–9731, 2024. 3

[74] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 1

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

[76] Zuo-Liang Zhu, Jian Yang, and Beibei Wang. Gaussian splatting with discretized sdf for relightable assets. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2025. 3

[77] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qixing Huang. Videomv: Consistent multi-view generation based on large video generative model, 2024. 3