

Hi3D: Pursuing High-Resolution Image-to-3D Generation with Video Diffusion Models

Haibo Yang*
 School of Computer Science
 Fudan University
 China
 yanghaibo.fdu@gmail.com

Yang Chen
 HiDream.ai Inc.
 China
 c1enyang@hidream.ai

Yingwei Pan
 HiDream.ai Inc.
 China
 pandy@hidream.ai

Ting Yao
 HiDream.ai Inc.
 China
 tiyao@hidream.ai

Zhineng Chen†
 School of Computer Science
 Fudan University
 China
 zhinchen@fudan.edu.cn

Chong-Wah Ngo
 Singapore Management University
 Singapore
 cwngo@smu.edu.sg

Tao Mei
 HiDream.ai Inc.
 China
 tmei@hidream.ai

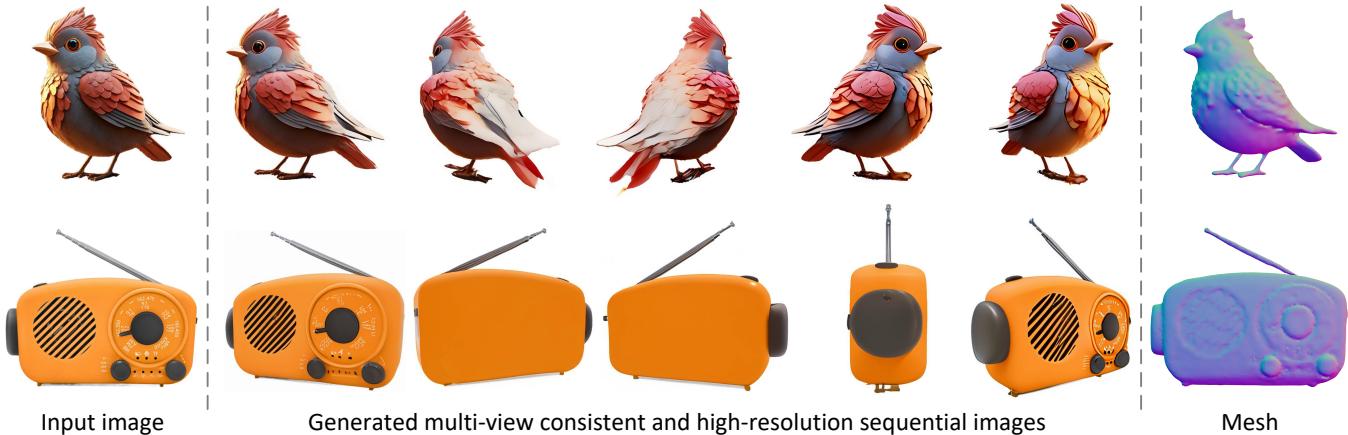


Figure 1: We propose *Hi3D*, the first high-resolution ($1,024 \times 1,024$) image-to-3D generation framework. Hi3D first generates multi-view consistent images from the input image and then reconstructs a high-fidelity 3D mesh from these generated images.

ABSTRACT

Despite having tremendous progress in image-to-3D generation, existing methods still struggle to produce multi-view consistent images with high-resolution textures in detail, especially in the paradigm of 2D diffusion that lacks 3D awareness. In this work, we present High-resolution Image-to-3D model (Hi3D), a new video

*This work was performed when Haibo Yang was visiting HiDream.ai as a research intern.

†Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0686-8/24/10
<https://doi.org/10.1145/3664647.3681634>

diffusion based paradigm that redefines a single image to multi-view images as 3D-aware sequential image generation (i.e., orbital video generation). This methodology delves into the underlying temporal consistency knowledge in video diffusion model that generalizes well to geometry consistency across multiple views in 3D generation. Technically, Hi3D first empowers the pre-trained video diffusion model with 3D-aware prior (camera pose condition), yielding multi-view images with low-resolution texture details. A 3D-aware video-to-video refiner is learnt to further scale up the multi-view images with high-resolution texture details. Such high-resolution multi-view images are further augmented with novel views through 3D Gaussian Splatting, which are finally leveraged to obtain high-fidelity meshes via 3D reconstruction. Extensive experiments on both novel view synthesis and single view reconstruction demonstrate that our Hi3D manages to produce superior multi-view consistency images with highly-detailed textures. Source code and data are available at <https://github.com/yanghb22-fdu/Hi3D-Official>.

CCS CONCEPTS

- Information systems → Multimedia content creation.

KEYWORDS

Image-to-3D generation; Video diffusion model; High resolution

ACM Reference Format:

Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Chong-Wah Ngo, and Tao Mei. 2024. Hi3D: Pursuing High-Resolution Image-to-3D Generation with Video Diffusion Models. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3664647.3681634>

1 INTRODUCTION

Image-to-3D generation, i.e., the task of reconstructing 3D mesh of object with corresponding texture from only a single-view image, has been a fundamental problem in multimedia [7, 8, 37] and computer vision [41, 68] fields for decades. In the early stage, the typical solution is to capitalize on regression or retrieval approaches [24, 55] for 3D reconstruction, which tends to be confined to close-world data with category-specific priors. This direction inevitably fails to scale up in real-world data. Recently, the success of diffusion models [18, 19, 69] has led to widespread dominance for open-world image content creation [34, 39, 44, 46, 47, 50]. Inspired by this, modern image-to-3D studies turn the focus on exploring how to exploit 2D prior knowledge from the pre-trained 2D diffusion model for image-to-3D generation in a two-phase manner, i.e., first multi-view images generation and then 3D reconstruction. One representative practice Zero123 [26] remoulds the text-to-image 2D diffusion model for viewpoint-conditioned image translation, which exhibits promising zero-shot generalization capability for novel view synthesis. Nevertheless, such independent modeling between the input image and each novel-view image might result in severe geometry inconsistency across multiple views. To alleviate this issue, several subsequent works [21, 27, 29, 48, 49, 53] further upgrade the 2D diffusion paradigm by simultaneously triggering image translation between the input image and multi-view images. Despite improving multi-view images generation, these approaches in 2D diffusion paradigm still suffer from multi-view inconsistency issues especially for complex object geometry. The underlying rationale is that the pre-trained 2D diffusion model is exclusively trained on individual 2D images, therefore lacking 3D awareness and resulting in sub-optimal multi-view consistency. Moreover, the geometry inconsistency among the output multi-view images will affect the overall stability of single-to-multi-view image translation during training. Hence, existing Image-to-3D techniques [21, 27, 29] mostly reduce the image size to low resolution (256×256). Such way practically increases batch size and improves training stability, while sacrificing the visual quality of output images. This severely hinders their applicability in many real-world scenarios that require high-fidelity 3D mesh with higher-resolution texture details, such as Virtual Reality and 3D film production.

In response to the above issues, our work paves a new way to formulate image translation across different views as 3D-aware sequential image generation (i.e., orbital video generation) by capitalizing on the pre-trained video diffusion model. Different from

2D diffusion model that lacks 3D awareness, video diffusion model is trained with a large volume of sequential frame images, and the learnt temporal consistency knowledge among frames can be naturally interpreted as one kind of 3D geometry consistency across multi-view images, especially for orbital videos. This motivates us to excavate such 3D prior knowledge from the pre-trained video diffusion model to enhance image-to-3D generation. More importantly, such video diffusion based paradigm enables more stable sequential image generation with amplified 3D geometry consistency. It in turn allows flexible scaling up of higher-resolution sequential image generation (e.g., $256 \times 256 \rightarrow 1,024 \times 1,024$), triggering 3D mesh generation with higher-resolution texture details.

By consolidating the idea of framing image-to-3D in video diffusion based paradigm, we novelly present High-resolution Image-to-3D model (Hi3D), to facilitate the generation of multi-view consistent meshes with high-resolution detailed textures in two-stage manner. Specifically, in the first stage, a pre-trained video diffusion model is remoulded with additional condition of camera pose, targeting for transforming single-view image into low-resolution 3D-aware sequential images (i.e., orbit video with 512×512 resolution). In the second stage, this low-resolution orbit video is further fed into 3D-aware video-to-video refiner with additional depth condition, leading to high-resolution orbit video ($1,024 \times 1,024$) with highly detailed texture. Considering that the obtained high-resolution orbit video contains a fixed number of multi-view images, we augment them with more novel views through 3D Gaussian Splatting. The resultant dense high-resolution sequential images effectively ease the final 3D reconstruction, yielding high-quality 3D meshes.

The main contribution of this work is the proposal of the two-stage video diffusion based paradigm that fully unleashes the power of inherent 3D prior knowledge in the pre-trained video diffusion model to strengthen image-to-3D generation. This also leads to the elegant views of how video diffusion model should be designed for fully exploiting 3D geometry priors, and how to scale up the resolution of multi-view images for high-resolution image-to-3D generation. Extensive experiments demonstrate the state-of-the-art performances of our Hi3D on both novel view synthesis and single view reconstruction tasks.

2 RELATED WORKS

Image-to-3D generation. Recently, with the remarkable advances in text-to-image diffusion models [18, 19], image-to-3D generation has also gained significant progress. These works can be generally categorized into three groups. The first group is optimize-based approaches [31, 40, 43, 54, 61]. Motivated by the pioneering work DreamFusion [38] in text-to-3D generation [4–6, 62, 63], this direction focuses on per-scene optimization by leveraging the prior knowledge in the pre-trained 2D diffusion model through score distillation sampling. While these methods have shown promising results, they often require extensive optimization time. To overcome this issue, the second group explores the direct training of image conditional 3D generative models [3, 10, 22, 28, 35, 65, 66]. Nonetheless, the limited availability of diverse 3D data has hampered these models' ability to generalize, with many studies being validated only on a narrow range of shape categories. The third direction is the recently emerging two-stage approach [21, 27, 29, 48, 49, 53],

which first generates multi-view images, and then reconstructs the corresponding 3D model. These methods achieve impressive results and have a fast generation speed. Our work also falls into this group. However, unlike previous methods that capitalize on the 2D diffusion model, we remodel the video diffusion model for 3D-aware multi-view image generation. This can fully unleash the power of inherent 3D prior knowledge in the pre-trained video diffusion model to strengthen image-to-3D generation. We note that some concurrent works [9, 17, 56] also use video diffusion models for 3D generation. The key difference is that we capitalize on video diffusion model to devise a novel 3D-aware video-to-video refiner, which not only scales up the resolution of the generated multi-view images but also refines 3D details & consistency.

3D Reconstruction. The recent success of neural radiance fields (NeRFs) [32] has inspired many follow-up works [15, 33, 57] to achieve impressive 3D reconstruction. However, these methods typically necessitate over a hundred images for training views, and their efficacy in reconstructing 3D models from sparse multi-view images remains suboptimal. To address this issue, several studies have endeavored to minimize the requisite number of training views. For instance, DS-NeRF [13] introduced additional depth supervision to enhance rendering quality, while RegNeRF [36] developed a depth smoothness loss for geometric regularization to facilitate training stability. Sparseneus [30] focused on learning geometry encoding priors from image features for adaptable neural surface learning from sparse input views, though the detail in reconstruction results was still lacking. In this work, we develop a straightforward yet efficient reconstruction pipeline that leverages the state-of-the-art 3D Gaussian Splatting algorithm [23] to augment the generated multi-view images, which enables us to stably and effectively reconstruct high-quality meshes.

3 PRELIMINARIES

Video Diffusion Models. Diffusion models [18, 51] are generative models that can learn the target data distribution from a Gaussian distribution through a gradual denoising process. Video diffusion models [2, 20, 60] are usually built upon pre-trained image diffusion models [34, 46], and enable the denoising process over multiple frames simultaneously. For simplicity, we adopt Stable Video Diffusion [1] as the basic video diffusion model, which achieves state-of-the-art performance in image-to-video generation. Formally, given a single frame x^0 , video diffusion model can generate a high-fidelity video consisting of N sequential frames $\mathbf{x} = \{x^0, x^1, \dots, x^{(N-1)}\}$ through an iterative denoising process. Specifically, at each denoising step t , video diffusion model predicts the amount of noise added in the sequence through a conditional 3D-UNet Φ , and then denoises the sequence by subtracting the predicted noise:

$$\mathbf{x}_{t-1} = \Phi(\mathbf{x}_t; t, c), \quad (1)$$

where c is the condition embedding of the input frame. In practice, Stable Video Diffusion is built within a latent diffusion framework [46] to reduce computational complexity, i.e., operating diffusion process in an encoded latent space. In this way, the input video sequence is first encoded into a latent code by a pre-trained VAE encoder and the denoised latent code is decoded back to pixel space using a VAE decoder after the denoising steps. Note that Stable

Video Diffusion is pre-trained on large-scale high-quality video datasets and demonstrates impressive image-to-video generation capacity. In this work, we propose to inherit the underlying temporal consistency knowledge in video diffusion model to boost the multi-view consistency for image-to-3D generation.

3D Gaussian Splatting. 3D Gaussian Splatting (3DGS) [23] emerges as a recent groundbreaking technique for novel view synthesis. Unlike 3D implicit representation methods (e.g., Neural Radiance Fields (NeRF) [32]) that rely on computationally intensive volume rendering for image generation, 3DGS achieves real-time rendering speeds through a splatting approach [64]. Specifically, 3DGS represents a 3D scene as a set of scaled 3D Gaussian primitives, and each scaled 3D Gaussian G_k is parameterized by an opacity (scale) $\alpha_k \in [0, 1]$, view-dependent color $c_k \in \mathbb{R}^3$, center position $\mu_k \in \mathbb{R}^{3 \times 1}$, covariance matrix $\Sigma_k \in \mathbb{R}^{3 \times 3}$. The 3D Gaussians can be queried as follows:

$$G_k(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \Sigma_k^{-1} (\mathbf{x}-\mu_k)}. \quad (2)$$

3DGS computes the color of each pixel via alpha blending according to the primitive's depth order $1, \dots, K$:

$$C(\mathbf{x}) = \sum_{k=1}^K c_k \sigma_k \prod_{j=1}^{k-1} (1 - \sigma_j), \sigma_k = \alpha_k G_k(\mathbf{x}). \quad (3)$$

Since the rendering process in 3DGS is fast and differentiable, the parameters of 3D Gaussian can be efficiently optimized through a multi-view loss (see [23] for more details). In this paper, we integrate 3DGS into our 3D reconstruction pipeline to extract high-fidelity meshes, tailored for synthesized high-resolution multi-view images.

4 OUR APPROACH

In this work, we devise a new High-resolution image-to-3D generation architecture, namely Hi3D, to novelly integrate video diffusion models into 3D-aware 360° sequential image generation (i.e., orbital video generation). Our launching point is to exploit the intrinsic temporal consistent knowledge in video diffusion models to enhance cross-view consistency in 3D generation. We begin this section by elaborating the problem formulation of image-to-3D generation (Sec. 4.1). We then elaborate the details of two-stage video diffusion based paradigm in our Hi3D framework. Specifically, in the first stage, we remould the pre-trained image-to-video diffusion model with additional condition of camera pose and then fine-tune it on 3D data to enable orbital video generation (Sec. 4.2). In the second stage, we further scale up the multi-view image resolution through a 3D-aware video-to-video refiner (Sec. 4.3). Finally, a novel 3D reconstruction pipeline is introduced to extract high-quality 3D mesh from these high-resolution multi-view images (Sec. 4.4). The whole architecture of Hi3D is illustrated in Figure 2.

4.1 Problem Formulation

Given a single RGB image $I \in \mathbb{R}^{3 \times H \times W}$ (source view) of an object X , our target is to generate its corresponding 3D content (i.e., textured triangle mesh). Similar to previous image-to-3D generation methods, we also decompose this challenging task into two steps: 1) generate a sequence of multi-view images around the object X and 2) reconstruct the 3D content from these generated multi-view images. Technically, we first synthesize a sequence of multi-view

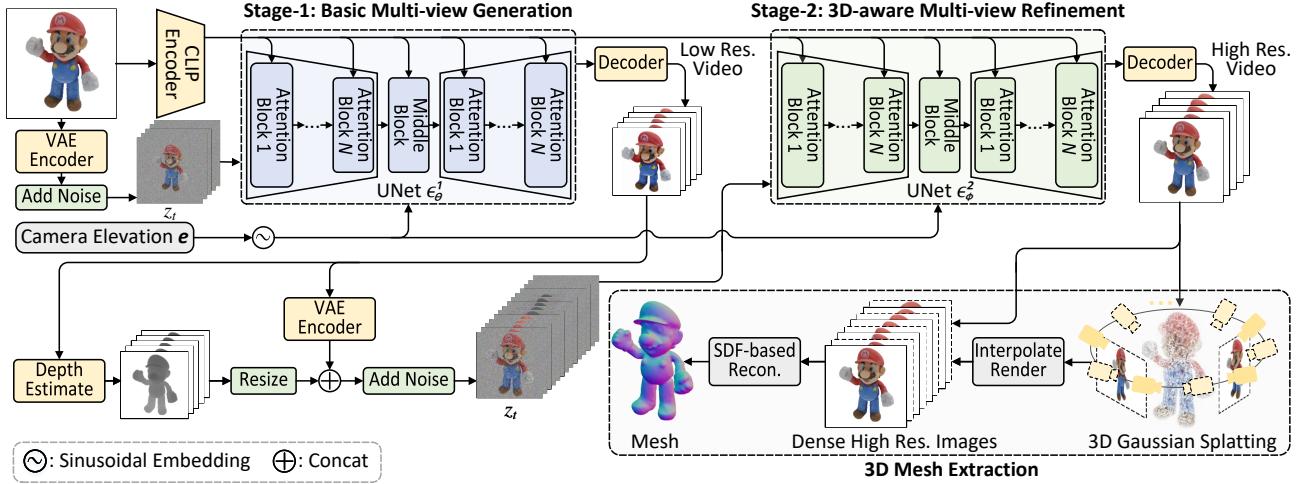


Figure 2: An overview of our proposed Hi3D. Our Hi3D fully exploits the capabilities of large-scale pre-trained video diffusion models to effectively trigger high-resolution image-to-3D generation. Specifically, in the first stage of basic multi-view generation, Hi3D remoulds video diffusion model with additional camera pose condition, aiming to transform single-view image into low-resolution 3D-aware sequential images. Next, in the second stage of 3D-aware multi-view refinement, we feed this low-resolution orbit video into 3D-aware video-to-video refiner with additional depth condition, leading to high-resolution orbit video with highly detailed texture. Finally, we augment the resultant multi-view images with more novel views through 3D Gaussian Splatting and employ SDF-based reconstruction to extract high-quality 3D meshes.

images $F \in \mathbb{R}^{N \times 3 \times H \times W}$ of the object from N different camera poses $\pi \in \mathbb{R}^{N \times 3 \times 4}$ corresponding to the input condition image I in a two-stage manner. Herein, we generate $N = 16$ multi-view images with a high resolution of $H \times W = 1,024 \times 1,024$ around the object in this work. It is worth noting that previous state-of-the-art image-to-3D models [21, 27, 29] can only generate low-resolution (i.e., 256×256) multi-view images. In contrast, to the best of our knowledge, our work is the first to enable high-resolution (i.e., $1,024 \times 1,024$) image-to-3D generation, which can preserve richer geometry and texture details of the input image. Next, we extract 3D mesh from these synthesized high-resolution multi-view images through our carefully designed 3D reconstruction pipeline. Since the number of generated views is somewhat limited, it is difficult to extract a high-quality mesh from these sparse views. To alleviate this issue, we leverage the novel view synthesis method (3D Gaussian Splatting [23]) to reconstruct an implicit 3D model from multi-view images F . Then we render additional interpolation views $F^* \in \mathbb{R}^{M \times 3 \times H \times W}$ between the multi-view images and add these rendered views into F , thereby obtaining dense view images $K \in \mathbb{R}^{(N+M) \times 3 \times H \times W} = F + F^*$ of the object X . Finally, we adopt an SDF-based reconstruction method [57] to extract a high-quality mesh from these dense views K .

4.2 Stage-1: Basic Multi-view Generation

Previous image-to-3D generation methods [21, 27, 29, 48] usually rely on pre-trained image diffusion models to accomplish multi-view generation. These methods generally extend the 2D UNet in image diffusion models to 3D UNet by injecting multi-view cross-attention layers. These added attention layers are trained from scratch on 3D datasets to learn multi-view consistency. However,

the image resolution in these methods is restricted to 256×256 to ensure training stability. Maintaining the original resolution (512×512) in pre-trained image diffusion models will lead to slower convergence and higher variance, as pointed in Zero123 [26]. Consequently, due to such low-resolution limitation, these methods fail to fully capture the primary rich 3D geometry and texture details in the input 2D image. In addition, we observe that these approaches still suffer from multi-view inconsistency issue, especially for complex object geometry. This may be attributed to the fact that the underlying pre-trained 2D diffusion model is exclusively trained on individual 2D images and lacks 3D modeling of multi-view correlation. To alleviate the above issues, we redefine single image to multi-view images as 3D-aware sequence image generation (i.e., orbital video generation) and utilize pre-trained video diffusion models to fulfill this goal. In particular, we repurpose Stable Video Diffusion (SVD) [1] to generate multi-view images from the input image. SVD is appealing because it was trained on a large variety of videos, which allows the network to encounter multiple views of an object during training. This potentially alleviates the 3D data scarcity problem. Moreover, SVD has already explicitly modeled the multi-frame relation via temporal attention layers. We can inherit the intrinsic multi-frame consistent knowledge in these temporal layers to pursue multi-view consistency in 3D generation.

Training Data. We first construct a high-resolution multi-view image dataset from the LVIS subset of the Objaverse [12]. For each 3D asset, we render 16 views with $1,024 \times 1,024$ resolution at random elevation $e \in [-10^\circ, 40^\circ]$. It is important to note that while the elevation is randomly selected, it remains the same across all views within a single video. For each video, the cameras are positioned equidistantly from the object with distance $r = 1.5$ and spaced evenly from 0° to 360° in azimuth angle. In total, our

training dataset comprises approximately 300,000 videos, denoted as $\mathcal{J} = \{\mathbf{J}_i, \mathbf{I}_i, e_i\}$, where the input condition image $\mathbf{I}_i = [\mathbf{J}_i]_1$ is the first frame in sequential images \mathbf{J}_i .

Video Diffusion Fine-tuning. In the first stage, our goal is to repurpose the pre-trained image-to-video diffusion model to generate multi-view consistent sequential images. The aforementioned multi-view image dataset $\mathcal{J} = \{\mathbf{J}_i, \mathbf{I}_i, e_i\}$ is thus leveraged to fine-tune the 3D-aware video diffusion model with additional camera pose condition. Specifically, given the input single-view image \mathbf{I}_i , we first project it into latent space by the VAE encoder of video diffusion model, and channel-wisely concatenate it with the noisy latent sequence, which encourages synthesized multi-view images to preserve the identity and intricate details of the input image. In addition, we incorporate the input condition image's CLIP embeddings [42] into the diffusion UNet through cross-attention mechanism. Within each transformer block, the CLIP embedding matrix acts as the key and value for the cross-attention layers, coupled with the layer's features serving as the query. In this way, the high-level semantic information of the input image is propagated into the video diffusion model. Since the multi-view image sequence is rendered at random elevations, we send the elevation parameter into the video diffusion model as additional condition. Most specifically, the camera elevation angle e is first embedded into sinusoidal positional embeddings and then fed into the UNet along with the diffusion noise timestep t . As all multi-view sequences follow the same azimuth trajectory, we do not send the azimuth parameter into the diffusion model. Herein, we omit the original “fps id” and “motion bucket id” conditions in video diffusion model as these conditions are irrelevant to multi-view image generation.

In general, the denoising neural network (3D UNet) in our remolded video diffusion model can be represented as $\epsilon_\theta^1(\mathbf{z}_t; \mathbf{I}, t, e)$. Given the multi-view image sequence \mathbf{J} , the pre-trained VAE encoder $\mathcal{E}(\cdot)$ first extracts the latent code of each image to constitute a latent code sequence \mathbf{z} . Next, Gaussian noise $\epsilon \sim N(0, I)$ is added to \mathbf{z} through a typical forward diffusion procedure at each time step t to get the noise latent code \mathbf{z}_t . The 3D UNet $\epsilon_\theta^1(\mathbf{z}_t; \mathbf{I}, t, e)$ with parameter θ is trained to estimate the added noise ϵ based on the noisy latent code \mathbf{z}_t , input image condition \mathbf{I} and elevation angle e through the standard mean square error (MSE) loss:

$$\mathcal{L}_{Stage-1} = \mathbb{E}_{\mathbf{I}, \mathbf{J}, e, t, \epsilon} \left[\|w(t)(\epsilon_\theta^1(\mathbf{z}_t; \mathbf{I}, e, t) - \epsilon)\|_2^2 \right], \quad (4)$$

where $w(t)$ is a corresponding weighing factor.

Instead of directly training denoising neural network in high resolution (i.e., $1,024 \times 1,024$), we decompose this non-trivial problem into more stable sub-problems in a coarse-to-fine manner. In the first stage, we train the denoising neural network by using Eq. (4) with 512×512 resolution for low-resolution multi-view image generation. The second stage further transforms 512×512 multi-view images into high-resolution ($1,024 \times 1,024$) multi-view images.

4.3 Stage-2: 3D-aware Multi-view Refinement

The output 512×512 multi-view images of Stage-1 exhibit promising multi-view consistency, while still failing to fully capture the geometry and texture details of inputs. To address this issue, we include an additional stage to further scale up the low-resolution outputs of the first stage through a new 3D-aware video-to-video

refiner, leading to higher-resolution (i.e., $1,024 \times 1,024$) multi-view images with finer 3D details and consistency.

In this stage, we also remould the pre-trained video diffusion model as 3D-aware video-to-video refiner. Formally, such denoising neural network can be formulated as $\epsilon_\phi^2(\mathbf{z}_t; \mathbf{I}, \hat{\mathbf{J}}, \mathbf{D}, t, e)$, where $\hat{\mathbf{J}}$ denotes the generated multi-view images corresponding the input image \mathbf{I} in Stage-1, \mathbf{D} is the estimated depth sequence of the generated multi-view images $\hat{\mathbf{J}}$. To be clear, the input conditions \mathbf{I} and e are injected into pre-trained video diffusion model by the same way as in Stage-1. Besides, we adopt the VAE encoder to extract the latent code sequence of the pre-generated multi-view images $\hat{\mathbf{J}}$ and channel-wisely concatenate them with the noise latent \mathbf{z}_t as conditions. Moreover, to fully exploit the underlying geometry information of the generated multi-view images, we leverage an off-the-shelf depth estimation model [45] to estimate the depth of each image in $\hat{\mathbf{J}}$ as 3D cues, yielding a depth map sequence \mathbf{D} . We then directly resize the depth maps into the same resolution of the latent code \mathbf{z}_t , and channel-wisely concatenate them with \mathbf{z}_t . Finally, the remoulded denoising neural network is trained through standard MSE loss in diffusion models:

$$\mathcal{L}_{Stage-2} = \mathbb{E}_{\mathbf{I}, \mathbf{J}, \mathbf{D}, e, t, \epsilon} \left[\|w(t)(\epsilon_\phi^2(\mathbf{z}_t; \mathbf{I}, \hat{\mathbf{J}}, \mathbf{D}, e, t) - \epsilon)\|_2^2 \right], \quad (5)$$

where $w(t)$ is a weighing factor. Note that the resolution of training images in Eq. (5) is scaled up to $1,024 \times 1,024$.

During training, we adopt some image degradation methods [58] to synthesize $\hat{\mathbf{J}}$ for data augmentation, instead of solely using the generated coarse multi-view images from Stage-1. In particular, we utilize a high-order degradation model to synthesize training data, including a series of blur, resize, noise, and compression processes. To replicate overshoot artifacts (e.g., ringing or ghosting around sharp transitions in images), we utilize *sinc* filter. Additionally, random masking techniques are used to simulate the effect of shape deformation. This way not only accelerates the training process, but also enhances the robustness of our video-to-video refiner.

4.4 3D Mesh Extraction

Through the above two-stage video diffusion based paradigm, we can obtain a high-resolution image sequence $\mathbf{F} \in \mathbb{R}^{N \times 3 \times H \times W}$ ($N = 16, H = W = 1,024$) conditioned on the input image \mathbf{I} . In this section, we aim to extract high-quality meshes from these generated high-resolution multi-view images. Previous image-to-3D methods [21, 27, 29] usually reconstruct the target 3D mesh from the output image sequence by optimizing the neural implicit Signed Distance Field (SDF) [16, 57]. Nevertheless, these SDF-based reconstruction methods are originally tailored for dense image sequences captured in the real world, which commonly fail to reconstruct high-quality mesh based on only sparse views.

To alleviate this issue, we design a unique 3D reconstruction pipeline for high-resolution sparse views. Instead of directly adopting SDF-based reconstruction methods to extract 3D mesh, we first use the 3D Gaussian Splatting (3DGS) algorithm [23] to learn an implicit 3D model from the generated high-resolution image sequence. 3DGS has demonstrated remarkable novel view synthesis capabilities and impressive rendering speed. Herein we attempt to utilize 3DGS's implicit reconstruction ability to augment the output sparse multi-view images of Stage-2 with more novel views. Specifically,

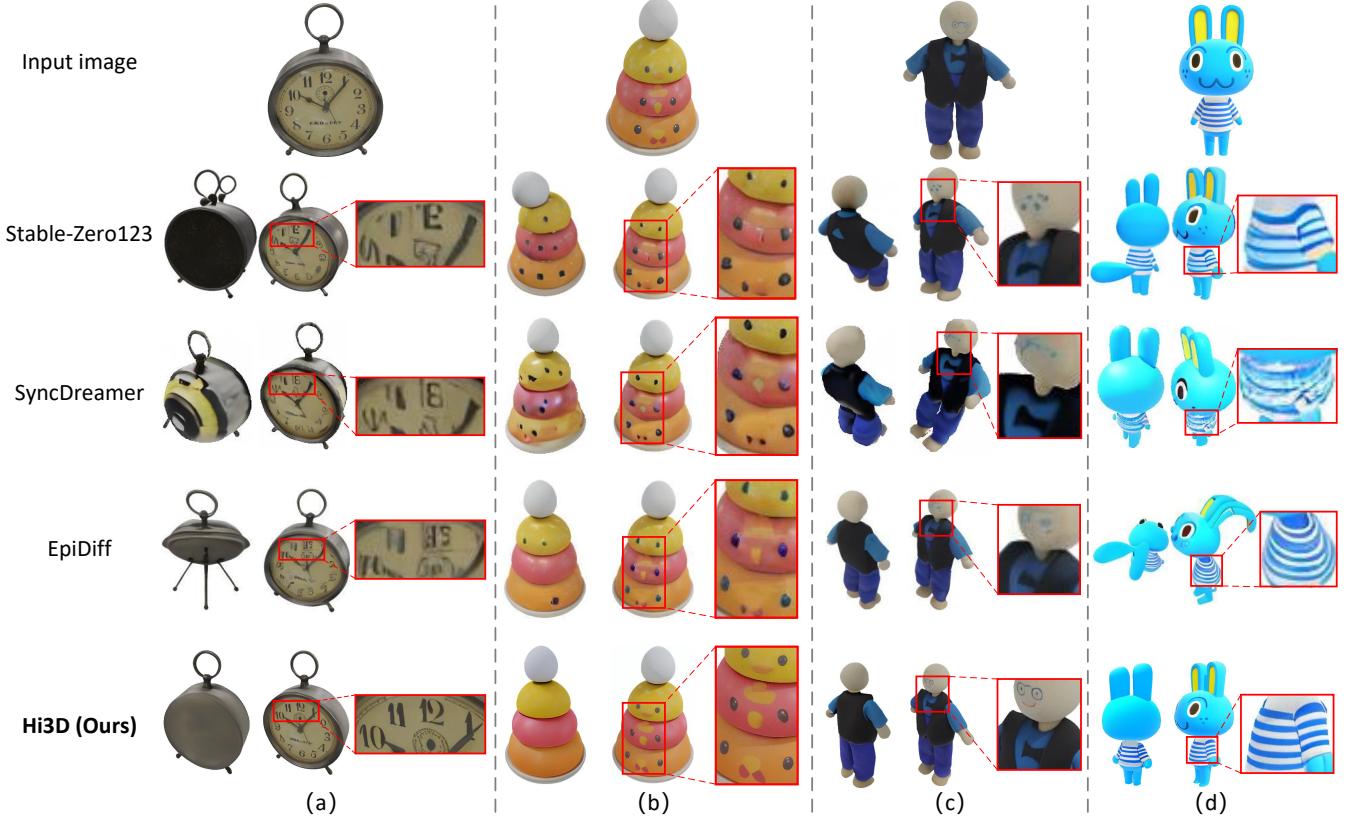


Figure 3: Qualitative comparisons with Stable-Zero123 [52], SyncDreamer [27] and EpiDiff [21] on novel view synthesis task. Our Hi3D generates high-resolution multi-view images with remarkable consistent details.

Table 1: Quantitative comparison with state-of-the-art methods in novel view synthesis on GSO dataset.

| Method | PSNR↑ | SSIM↑ | LPIPS↓ |
|---------------------|--------------|--------------|--------------|
| Realfusion [31] | 15.26 | 0.722 | 0.283 |
| Zero123 [26] | 18.93 | 0.779 | 0.166 |
| Zero123-XL [11] | 19.47 | 0.783 | 0.159 |
| Stable-Zero123 [52] | 19.79 | 0.788 | 0.153 |
| SyncDreamer [27] | 20.05 | 0.798 | 0.146 |
| EpiDiff [21] | 20.49 | 0.855 | 0.128 |
| Hi3D (Ours) | 24.26 | 0.864 | 0.119 |

we render M interpolation views \mathbf{F}^* between the adjacent images in \mathbf{F} from the reconstructed 3DGs. Finally, we optimize an SDF-based reconstruction method [57] based on the augmented dense views $\mathbf{F} + \mathbf{F}^*$ to extract the high-quality 3D mesh of the object X .

5 EXPERIMENTS

5.1 Experimental Settings

Datasets and Evaluation. We empirically validate the merit of our Hi3D model by conducting experiments on two primary tasks, i.e., novel view synthesis and single view reconstruction. Following [21, 27, 29], we perform quantitative evaluation on Google Scanned Object (GSO) dataset [14]. For novel view synthesis task, we employ three commonly adopted metrics: PSNR, SSIM [59], and LPIPS [67].

For the single view reconstruction task, we use Chamfer Distances and Volume IoU to measure the quality of the reconstructed 3D models. In addition, to assess the generalization ability of our Hi3D, we perform qualitative evaluation over single images with various styles derived from the internet.

Implementation Details. During the first stage of basic multi-view generation, we downscale the video dataset as 512×512 videos. For the second stage of multi-view refinement, we not only feed the outputs of the first stage, but also adopt synthetic data generation strategy (similar to traditional image/video restoration methods [58]) for data augmentation. This strategy aims to accelerate the training process and enhance the model’s robustness. The overall experiments are conducted on eight 80G A100 GPUs. Specifically, the first stage undergoes 80,000 training steps (approximately 3 days), with a learning rate of 1×10^{-5} and a total batch size of 16. The second stage contains 20,000 training steps (around 3 days), with a learning rate of 5×10^{-5} and a reduced batch size of 8.

Compared Methods. We compare our Hi3D with the following state-of-the-art methods: RealFusion [31] and Magic123 [40] exploit 2D diffusion model (Stable Diffusion [46]) and SDS loss [38] for reconstructing from single-view image. Zero123 [26] learns to generate novel view images of the same object from different viewpoints, and can be integrated with SDS loss for 3D reconstruction. Zero123-XL [11] and Stable-Zero123 [52] further upgrade Zero123 by enhancing the training data quality. One-2-3-45 [25]

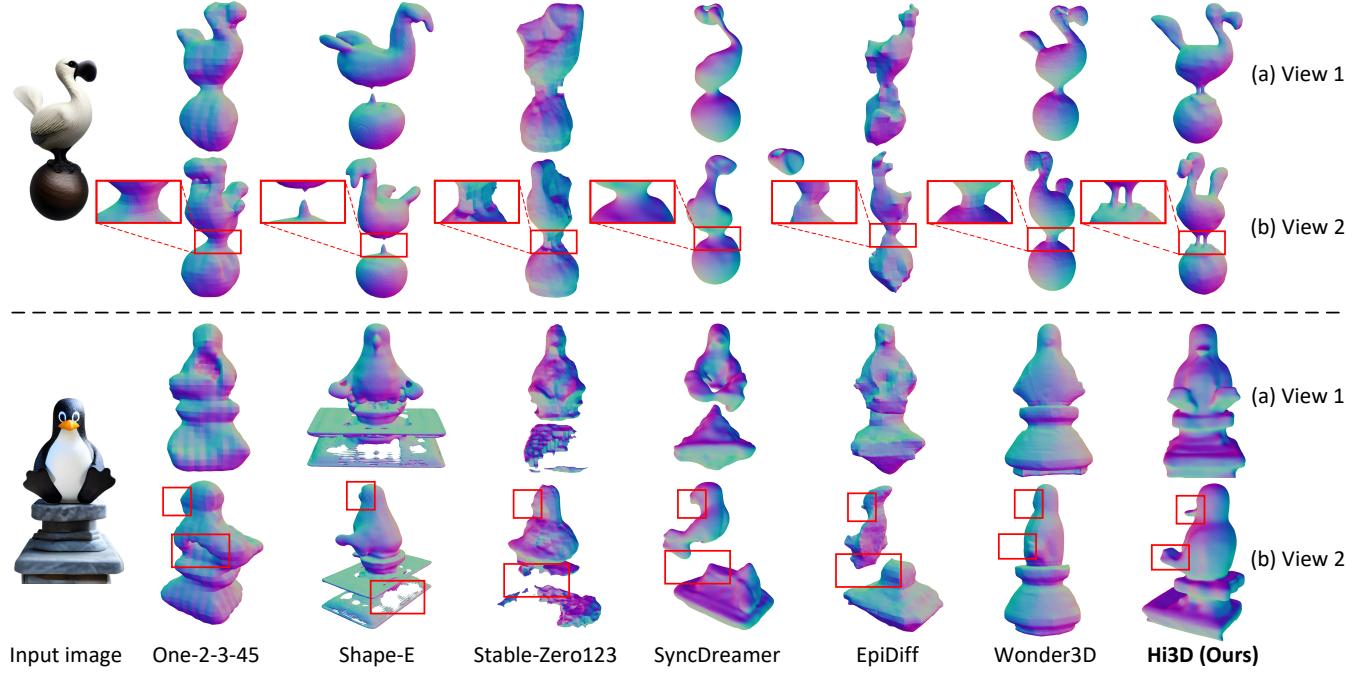


Figure 4: Qualitative comparison of 3D meshes generated by various methods on single view reconstruction task.

directly learns explicit 3D representation via 3D Signed Distance Functions (SDFs) [30] from multi-view images (i.e., the outputs of Zero123). Point-E [35] and Shape-E [22] are pre-trained over an extensive internal OpenAI 3D dataset, thereby being capable of directly transforming single-view images into 3D point clouds or shapes encoded in MLPs. SyncDreamer [27] introduces a 3D global feature volume to maintain multi-view consistency. Wonder3D [29] and EpiDiff [21] leverage 3D attention mechanisms to enable interaction among multi-view images via cross-attention layers. Note that in novel view synthesis task, we only include partial baselines (i.e., Zero123 series, SyncDreamer, EpiDiff) that can produce exactly the same viewpoints as our Hi3D for fair comparison.

5.2 Novel View Synthesis

Table 1 summarizes performance comparison on novel view synthesis task, and Figure 3 showcases qualitative results in two different views. Overall, our Hi3D consistently exhibits better performances than existing 2D diffusion based approaches. Specifically, Hi3D achieves the PSNR of 24.26%, which outperforms the best competitor EpiDiff by 3.77%. The highest image quality score of our Hi3D generally highlights the key advantage of video diffusion based paradigm that exploits 3D prior knowledge to boost novel view synthesis. In particular, due to the independent image translation, Zero123 series (e.g., Stable-Zero123) fails to achieve multi-view consistency results (e.g., one/two rings on the head of the alarm clock in different views in Figure 3 (a)). SyncDreamer and EpiDiff further strengthen multi-view consistency by exploiting 3D intermediate information or using multi-view attention mechanisms. Nevertheless, their novel-view results still suffer from blurry and unrealistic issues with degraded image quality (e.g., the blurry numbers of alarm clock in Figure 3 (a)) due to the restricted low image resolution

Table 2: Quantitative comparison with state-of-the-art methods in single view reconstruction on GSO dataset.

| Method | Chamfer Dist. \downarrow | Volume IoU \uparrow |
|---------------------|----------------------------|-----------------------|
| Realfusion [31] | 0.0819 | 0.2741 |
| Magic123 [40] | 0.0516 | 0.4528 |
| One-2-3-45 [25] | 0.0629 | 0.4086 |
| Point-E [35] | 0.0426 | 0.2875 |
| Shape-E [22] | 0.0436 | 0.3584 |
| Stable-Zero123 [52] | 0.0321 | 0.5207 |
| SyncDreamer [27] | 0.0261 | 0.5421 |
| EpiDiff [21] | 0.0343 | 0.4927 |
| Wonder3D [29] | 0.0199 | 0.6244 |
| Hi3D (Ours) | 0.0172 | 0.6631 |

(256×256). Instead, by mining 3D priors and scaling up multi-view image resolution via video diffusion model, our Hi3D manages to produce multi-view consistent and high-resolution 1,024×1,024 images, leading to highest image quality (e.g., the clearly visible numbers in alarm clock in Figure 3 (a)).

5.3 Single View Reconstruction

Next, we evaluate the single view reconstruction performance of our Hi3D in Table 2. In addition, Figure 4 shows qualitative comparison between Hi3D and existing methods. In general, our Hi3D outperforms state-of-the-art methods over both two metrics. Specifically, One-2-3-45 directly leverages multi-view outputs of Zero123 with sub-optimal 3D consistency for reconstruction, which commonly results in over-smooth meshes with fewer details. Stable-Zero123 further improves 3D consistency with higher-quality training data,

Table 3: Ablation study on 3D-aware multi-view refinement.

| Setting | PSNR↑ | SSIM↑ | LPIPS↓ |
|----------------|--------------|--------------|--------------|
| Hi3D | 24.26 | 0.864 | 0.119 |
| w/o refinement | 22.09 | 0.842 | 0.136 |
| w/o depth | 23.12 | 0.848 | 0.128 |

Table 4: Ablation study on 3D reconstruction pipeline.

| Setting | Chamfer Dist.↓ | Volume IoU↑ |
|----------|----------------|---------------|
| $M = 0$ | 0.0186 | 0.6375 |
| $M = 16$ | 0.0172 | 0.6631 |
| $M = 32$ | 0.0174 | 0.6598 |
| $M = 48$ | 0.0175 | 0.6607 |

while still suffering from missing or over-smooth meshes. Different from independent image translation in Zero123, SyncDreamer, EpiDiff, and Wonder3D exploit simultaneous multi-view image translation through 2D diffusion model, thereby leading to better 3D consistency. However, they struggle to reconstruct complex 3D meshes with rich details due to the limitation of low-resolution multi-view images. In contrast, our Hi3D fully unleashes the power of inherent 3D prior knowledge in pre-trained video diffusion model and scales up the multi-view images into higher resolution. Such design enables higher-quality 3D mesh reconstruction with richer fine-grained details (e.g., the feet of bird and penguin in Figure 4).

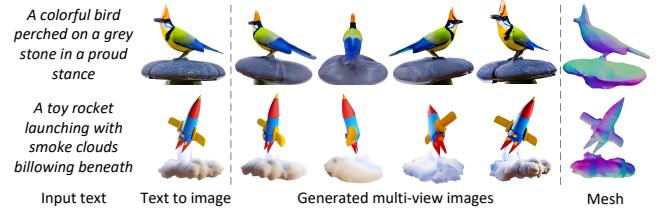
5.4 Ablation Studies

Effect of 3D-aware Multi-view Refinement Stage. Here we examine the effectiveness of the second stage (i.e., 3D-aware multi-view refinement) on novel view synthesis. Table 3 details the performances of ablated runs of our Hi3D. Specifically, the second row removes the whole second stage, and the performances drop by a large margin. This validates the merit of scaling up multi-view image resolution via 3D-aware video-to-video refiner. In addition, when only removing depth condition in second stage (row 3), a clear performance drop is attained, which demonstrates the effectiveness of depth condition that enhances 3D geometry consistency among multi-view images.

Effect of Interpolation view number M in 3D Reconstruction. Table 4 shows the single view reconstruction performances of using different numbers of interpolation views M . In the extreme case of $M = 0$, no interpolation view is employed, and the 3D reconstruction pipeline degenerates to typical SDF-based reconstruction. By increasing M as 16, the reconstruction performances are clearly improved, which basically shows the advantage of interpolation views via 3DGs. However, when further enlarging M , the performances slightly decrease. We speculate that this may be the result of unnecessary information across views repeat and error accumulating. In practice, M is generally set to 16.

5.5 More Discussions

Text-to-image-to-3D. By integrating advanced text-to-image models (e.g., Stable Diffusion [46], Imagen [47]) into our Hi3D, we are capable of generating 3D models directly from textual descriptions,

**Figure 5: Examples of using Hi3D for text-to-3D generation.****Figure 6: Diverse and creative results of our Hi3D with different seeds.**

as illustrated in Figure 5. Our approach manages to produce higher-fidelity 3D models with highly-detailed texture, which again highlights the merit of high-resolution multi-view image generation with 3D consistency.

Diversity and Creativity in 3D Model Generation. Here we examine the diversity and creativity of our Hi3D by using different random seeds. As shown in Figure 6, our Hi3D is able to generate diverse and plausible instances, each with distinct geometric structures or textures. This capability not only enhances the flexibility of 3D model creation but also significantly contributes to the exploration of creative possibilities in 3D design and visualization.

6 CONCLUSION

This paper explores inherent 3D prior knowledge in pre-trained video diffusion model for boosting image-to-3D generation. Particularly, we study the problem from a novel viewpoint of formulating single image to multi-view images as 3D-aware sequential image generation (i.e., orbital video generation). To materialize our idea, we have introduced Hi3D, which executes two-stage video diffusion based paradigm to trigger high-resolution image-to-3D generation. Technically, in the first stage of basic multi-view generation, a video diffusion model is remoulded with additional 3D condition of camera pose, targeting for transforming single image into low-resolution orbital video. In the second stage of 3D-aware multi-view refinement, a video-to-video refiner with depth condition is designed to scale up the low-resolution orbital video into high-resolution sequential images with rich texture details. The resulting high-resolution outputs are further augmented with interpolation views through 3D Gaussian Splatting, and SDF-based reconstruction is finally employed to achieve 3D meshes. Experiments conducted on both novel view synthesis and single view

reconstruction tasks validate the superiority of our proposal over state-of-the-art approaches.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China (No. 2022YFB3104703) and in part by the National Natural Science Foundation of China (No. 62172103).

REFERENCES

- [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127* (2023).
- [2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *CVPR*.
- [3] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhiuwen Tu, Lingjie Liu, and Hao Su. 2023. Single-Stage Diffusion NeRF: A Unified Approach to 3D Generation and Reconstruction. In *ICCV*.
- [4] Yang Chen, Jingwen Chen, Yingwei Pan, Xinmei Tian, and Tao Mei. 2023. 3D Creation at Your Fingertips: From Text or Image to 3D Assets. In *ACM MM*.
- [5] Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. 2023. Control3d: Towards controllable text-to-3d generation. In *ACM MM*.
- [6] Yang Chen, Yingwei Pan, Haibo Yang, Ting Yao, and Tao Mei. 2024. Vp3d: Unleashing 2d visual prompt for text-to-3d generation. In *CVPR*.
- [7] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. 2019. Animating Your Life: Real-Time Video-to-Animation Translation. In *ACM MM*.
- [8] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. 2019. Mocycle-gan: Unpaired video-to-video translation. In *ACM MM*.
- [9] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. 2024. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738* (2024).
- [10] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. 2023. SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In *CVPR*.
- [11] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. In *NeurIPS*.
- [12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *CVPR*.
- [13] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised NeRF: Fewer Views and Faster Training for Free. In *CVPR*.
- [14] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*.
- [15] Qiancheng Fu, Qingshan Xu, Yew-Soo Ong, and Wenbing Tao. 2022. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. In *NeurIPS*.
- [16] Yuan-Chen Guo. 2022. Instant Neural Surface Reconstruction. <https://github.com/bennyguo/instant-nsr-pl>.
- [17] Junlin Han, Filippos Kokkinos, and Philip Torr. 2024. Vfusion3d: Learning scalable 3d generative models from video diffusion models. *arXiv preprint arXiv:2403.12034* (2024).
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.
- [19] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. In *NeurIPS Workshop*.
- [20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. In *NeurIPS*.
- [21] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. 2024. EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion. In *CVPR*.
- [22] Heewoo Jun and Alex Nichol. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463* (2023).
- [23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *TOG* (2023).
- [24] Xuetong Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. 2020. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*.
- [25] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, and Hao Su. 2023. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. In *NeurIPS*.
- [26] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*.
- [27] Yuan Liu, Cheng Lin, Zijian Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2024. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *ICLR*.
- [28] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. 2023. MeshDiffusion: Score-based generative 3d mesh modeling. In *ICLR*.
- [29] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. In *CVPR*.
- [30] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*.
- [31] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*.
- [32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.
- [33] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *TOG* (2022).
- [34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *PMLR*.
- [35] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751* (2022).
- [36] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *CVPR*.
- [37] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To create what you tell: Generating videos from captions. In *ACM Multimedia*.
- [38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2023. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*.
- [39] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. 2024. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations. In *CVPR*.
- [40] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. 2024. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. In *ICLR*.
- [41] Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. 2024. Boosting Diffusion Models with Moving Average Sampling in Frequency Domain. In *CVPR*.
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- [43] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. In *ICCV*.
- [44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).
- [45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI* (2020).
- [46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
- [47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.
- [48] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. 2023. Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. *arXiv preprint arXiv:2310.15110* (2023).
- [49] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2024. MVDream: Multi-view Diffusion for 3D Generation. In *ICLR*.
- [50] Yan Shu, Weichao Zeng, Zhenhang Li, Fangmin Zhao, and Yu Zhou. 2024. Visual Text Meets Low-level Vision: A Comprehensive Survey on Visual Text Processing. *arXiv preprint arXiv:2402.03082* (2024).

- [51] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *ICLR*.
- [52] StabilityAI. 2023. Stable Zero123. <https://stability.ai/news/stable-zero123-3d-generation>.
- [53] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. 2023. Viewset Diffusion:(0-) Image-Conditioned 3D Generative Models from 2D Data. In *ICCV*.
- [54] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*.
- [55] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhiwen Li, Vladlen Koltun, and Thomas Brox. 2019. What do single-view 3d reconstruction networks learn?. In *CVPR*.
- [56] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforet, Robin Rombach, and Varun Jampani. 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008* (2024).
- [57] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *NeurIPS*.
- [58] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*.
- [59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *TIP* (2004).
- [60] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2023. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647* (2023).
- [61] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. 2023. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 views. In *CVPR*.
- [62] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhiheng Chen, and Tao Mei. 2023. 3dstyle-diffusion: Pursuing fine-grained text-driven 3d stylization with 2d diffusion models. In *ACM MM*.
- [63] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhiheng Chen, Zuxuan Wu, Yu-gang Jiang, and Tao Mei. 2024. DreamMesh: Jointly manipulating and texturing triangle meshes for text-to-3d generation. In *ECCV*.
- [64] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztïreli, and Olga Sorkine-Hornung. 2019. Differentiable Surface Splatting for Point-based Geometry Processing. *TOG* (2019).
- [65] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. In *NeurIPS*.
- [66] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 2023. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. In *SIGGRAPH*.
- [67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.
- [68] Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. 2024. TRIP: Temporal Residual Learning with Image Noise Prior for Image-to-Video Diffusion Models. In *CVPR*.
- [69] Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. 2024. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer. In *CVPR*.

APPENDIX

Recall that in Stage-2 (see Sec. 4.3), we devise a new 3D-aware video-to-video refiner to further scale up the low-resolution (512×512) outputs of Stage-1 to higher resolution ($1,024 \times 1,024$). An alternative solution is to use a super-resolution (SR) model to directly upscale the generated multi-view images in Stage-1 into $1,024 \times 1,024$ resolution. Here we adopt a typical SR method (Real-ESRGAN [58]) for comparison.

Figure 7 showcases the comparison results. The SR method can only eliminate the blurriness and produce sharp outputs, but fails to alleviate the geometry and appearance distortions in the input multi-view images. Taking Figure 7 (a) as an example, compared with the input image, the “hands” and “face” in the generated images of Stage-1 are distorted. The SR method Real-ESRGAN cannot correct these distortions as it was primarily trained to produce high-resolution images strictly consistent with the input. Thus these distortions

inevitably remained in the SR outputs. In contrast, our devised 3D-aware video-to-video refiner not only produces clear results with no blur, but also generates correct “hand” and “face” that are consistent with the input image. This comparison clearly demonstrates the effectiveness of our proposed 3D-aware video-to-video refiner for generating high-resolution ($1,024 \times 1,024$) multi-view images with finer 3D details and consistency.

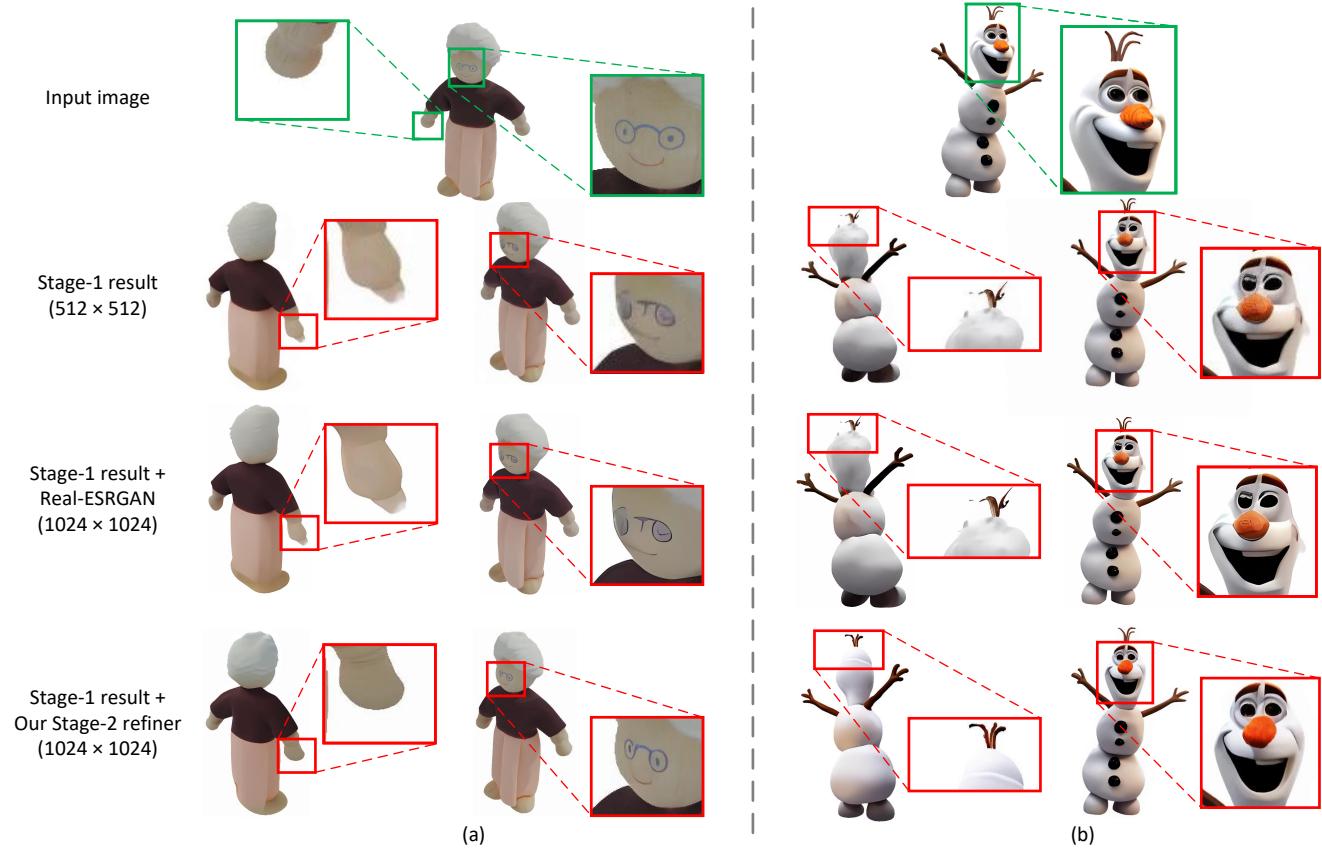


Figure 7: Comparing our 3D-aware video-to-video refiner with typical super-resolution method (Real-ESRGAN [58]) in Stage-2.