Unlocking the Video Prior for High-Fidelity Sparse Multi-View Image Synthesis

Fan Yang[1] Jianfeng Zhang[†2] Jun Hao Liew[2] Chaoyue Song[1]
Zhongcong Xu[2] Jiashi Feng[2] Guosheng Lin[†1]
[1]Nanyang Technological University [2]ByteDance Seed

Abstract

The development of multi-view image synthesis is constrained by the scarcity of training data. One promising solution is to finetune well-trained video generative models to synthesize 360-degree videos of objects. While these methods benefit from the strong generative priors inherited from the pretrained knowledge, they are limited by the high computational costs incurred by the large number of viewpoints. Existing methods commonly adopt temporal attention mechanism to address this. However, these methods suffer from undesirable artifacts such as 3D inconsistency and over-smoothing in the generated results. In this paper, we introduce a novel approach to unlock the video priors for multi-view synthesis by reducing generation into a sparser yet more precise process. Specifically, we introduce two strategies to achieve this: i) Condensing the video diffusion model to synthesize highly consistent sparse multi-view images. ii) Extracting dense geometrical priors from the pretrained video diffusion models to enhance the generation stability. The combination of these two strategies formulates a novel framework for multi-view synthesis, which is capable of synthesizing highly consistent sparse multi-view images with strong generalization ability. Extensive experiments demonstrate that our approach achieves superior efficiency, generalization, and consistency, outperforming state-of-the-art multi-view synthesis methods.

1. Introduction

The rise of video diffusion models has introduced new paradigms for novel view synthesis in 3D generation. Existing methods [6, 33, 45, 47] adapt multi-view synthesis into 360-degree video generation by fine-tuning pretrained video diffusion models [1, 34] on 3D-rendered datasets. The resulting multi-view video diffusion models offer two attractive advantages: i) strong generalization across diverse input cases, and ii) an intrinsic 3D structural constrain on the generative process that leads to better geometry quality. However, most of these methods rely on temporal attention to enforce the 3D consistency, which suffers from a limited receptive field over different views, often results in content drifting and over-smoothing, particularly under large camera movements.

A straightforward solution is to adopt denser attention mechanisms, such as 3D attention [23, 31], to replace temporal attention. However, this approach incurs prohibitively high computational costs due to the large number of generated views in video diffusion models. On the other hand, recent advances in large sparse-view reconstruction models, such as GRM [43], have focused on generating high-fidelity sparse multi-view images, which can be directly lifted into high-quality 3D assets. Despite the efficiency and practicality of this process, most existing methods for sparse multi-view generation [16, 23, 31] are fine-tuned from 2D diffusion models, leading to limited consistency and generalization capability.

In this work, we draw inspiration from these two techniques and introduce a novel approach to unlock video priors for high-fidelity multi-view synthesis by reducing the overall generation process into a sparser yet more precise pipeline. Specifically, we investigate this reduction from two perspectives: i) Condensing the pretrained video diffusion model into a sparser generation setting to improve the 3D consistency and generation efficiency. We adopt a denser attention mechanism to enforce 3D consistency and mitigate the associated computational overhead by reducing the number of synthesized views. During finetuning, we propose a novel distribution shift strategy that better preserves pretrained knowledge, leading to the generation of high-quality sparse multi-view images. ii) Extracting dense geometrical priors from the pretrained video diffusion models to further enhance the generation stability. Although finetuning the video diffusion models into a sparser one with a denser attention mechanism achieves relatively good results, we found that the pretrained knowledge established with the temporal continuity assumption in the pretrained video diffusion model may degrade as the number of views decreases, leading to unstable generation results for some in-the-wild cases. To address this problem, we

[†] Corresponding authors.

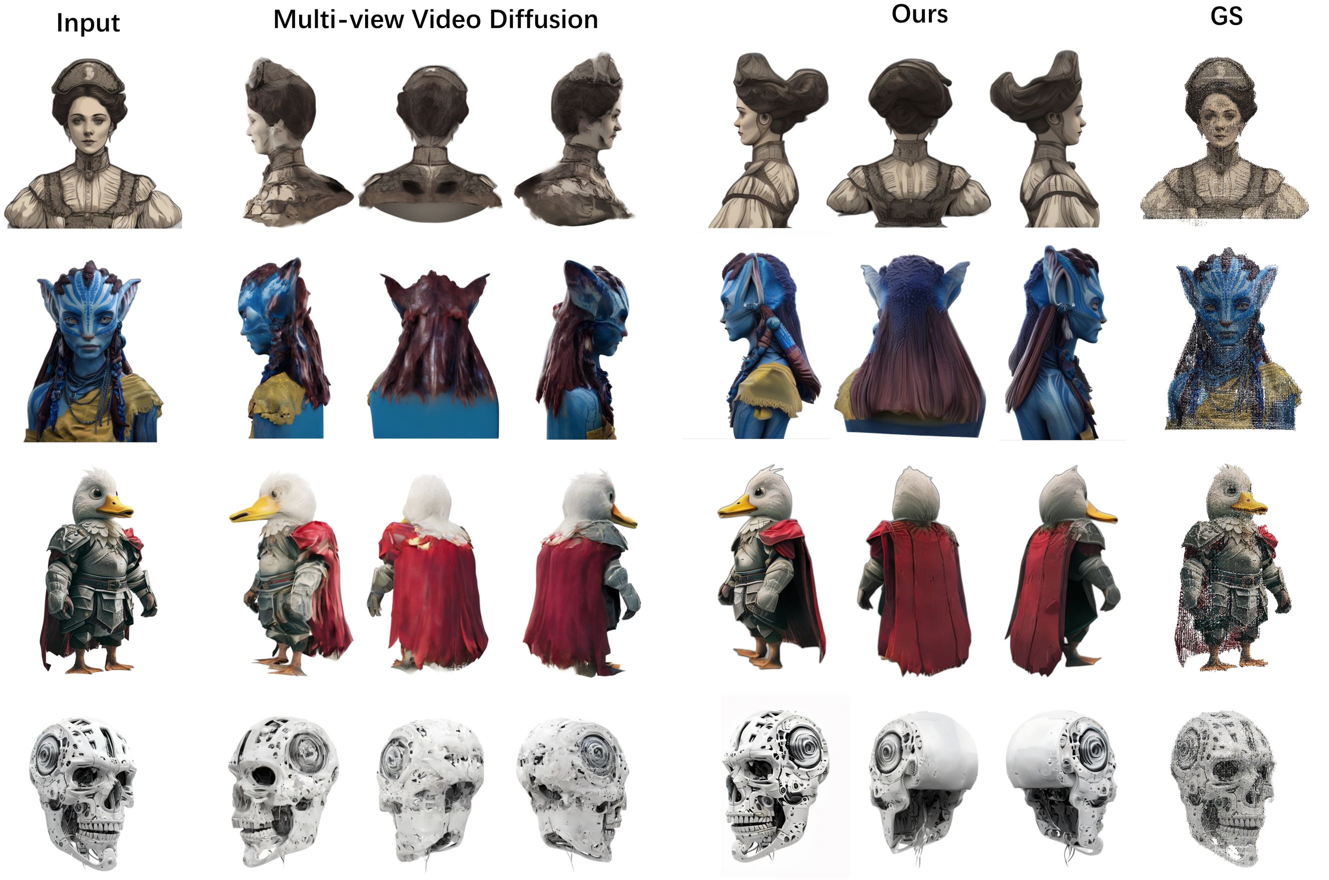

Figure 1. Comparison between our methods and multi-view video diffusion model SV3d(p) [33]. Our model excels at synthesizing highly consistent multi-view images while preserving the generative priors of video diffusion models. Although we adopt a significantly denser attention mechanism, the associated computational overhead is effectively mitigated by reducing the generation to sparse multi-view synthesis. This results in a 2× speedup during inference compared to SV3d(p), without compromising performance. From left to right, we show the input images, multi-view images synthesized by SV3d(p), our method and the reconstructed Gaussian Splatting from our synthesized outputs.

propose to extract dense geometrical priors from pretrained video diffusion models to guide the sparse multi-view generation process, further enhancing stability.

As shown in Figure 1, compared to existing multi-view video diffusion methods [33] that struggle to produce consistent results with temporal attention, our model excels at synthesizing highly consistent multi-view images while preserving the generative video priors. Extensive experiments on multiple datasets demonstrate that our model generates highly consistent novel views with superior 3D consistency and significantly improved generation efficiency. Our contributions could be summarized as follows: i) We introduce a novel approach to unlock video priors for high-fidelity sparse multi-view synthesis by reducing the generation process into a sparser yet more precise pipeline. ii) We propose to extract geometrical priors from multi-view video diffusion models to enhance the stability and quality of sparse multi-view generation. iii) Our approach improves both generation efficiency and 3D consistency, making video diffusion models more practical and effective for multi-view synthesis.

2. Related work

2.1. 3D Generation

3D generation has been extensively explored using various 3D representations including meshes [11, 22], voxels [2, 46], point clouds [4, 5], SDF [7, 25, 26], and Triplane [3, 12]. Traditional methods [13, 24], often trained on limited-scale 3D datasets, struggle to generate intricate geometric structures with substantial diversity.

The rise of diffusion models has opened new paradigms for 3D generation. Several methods are proposed to distill 3D information from large pretrained diffusion mod-

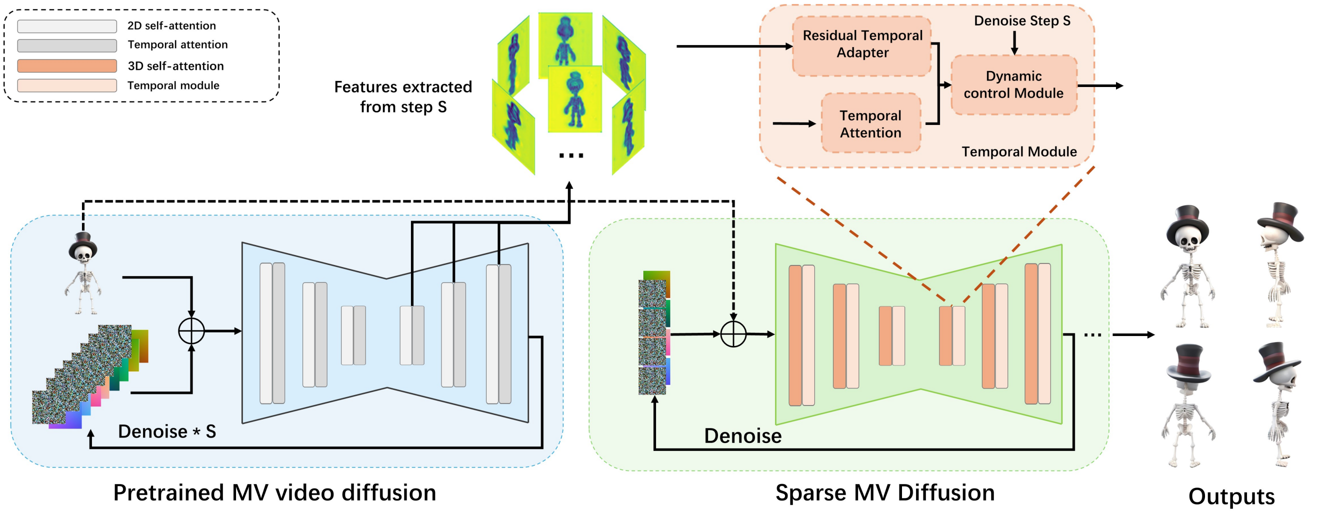

Figure 2. The overall framework of our model. We condense pretrained multi-view video diffusion model into a sparser yet more precise process by extending the constrained temporal attention with dense 3D attention and reducing the number of views to a small set (e.g., four). In addition, we propose to adopt a pretrained video diffusion model as a geometrical reference network to further enhance the generation process. The extracted video priors (from step S) are integrated into the condensed video diffusion network with the proposed Residual Temporal Adapter, which could be seamlessly plugged into the original network structure to guide the overall generation process.

els, which provide strong generative priors learned from the massive datasets. Notably, Score Distillation Sampling (SDS)-based methods [18, 27, 28, 36, 44] treat generation as an optimization problem, leveraging pretrained 2D diffusion models to supervise the unseen views of the target object. While these methods can generate realistic results, they suffer from slow convergence and the Janus problem, due to the lack of 3D understanding in the pretrained 2D diffusion models.

An alternative promising approach is to first generate multi-view images and then reconstruct the 3D shapes using techniques such as NeRF, Gaussian Splatting or feed-forward large reconstruction networks [17, 32, 38, 42, 43]. Although these methods achieve promising results, they still suffer from local inconsistencies, limited resolution of the input multi-view images, and are unable to generated 3D objects with complicated geometry and realistic textures.

2.2. Novel View Synthesis

The success of diffusion models has opened possibilities for novel view synthesis. Zero123 [20, 30] fine-tune a pretrained 2D diffusion model under different camera conditions to enable arbitrary view conditioned generation. Sparse novel view synthesis methods, such as MVDream [31], extend the original 2D self-attention to multi-view attention, achieving 3D-consistent multi-view image generation. Wonder3D [23] fine-tunes the 2D diffusion model with cross-domain RGB-normal attention layers to enhance the learning of geometry information and improve 3D consistency in the generated outputs.

However, due to the limited generalization ability of 2D diffusion models, these methods struggle to generate satisfactory results for out-of-domain inputs with complex geometry or textures. On the other hand, video diffusion models [1] have shown potential in rich generative priors for novel view synthesis [6, 41, 47]. SV3d [33] fine-tunes a pretrained video diffusion model on 3D-rendered datasets to synthesize 360-degree orbit videos. While these methods demonstrate great generalization, they still suffer from the limited receptive field of temporal attention, preventing them from generating highly consistent novel views, especially with large camera movement.

3. Method

Given a single-view image and target camera poses as inputs, our goal is to synthesize 3D-consistent sparse multi-view images that could be used to reconstruct 3D objects. To address this problem, we propose a novel framework that reduces a multi-view video diffusion model into a sparse generation setting, enabling the synthesis of high-fidelity multi-view images while largely preserving the pretrained generative priors.

The overall framework of our model is illustrated in Figure 2, comprising a *reduced multi-view generator* and a *geometrical reference network*. In contrast to prior works [16, 23, 31, 35, 39], which finetune pretrained 2D diffusion models to synthesize multi-view images, our approach condenses the pretrained multi-view video diffusion model into the setting of sparse multi-view synthesis, which enables the generator to inherit capabilities from the pretrained video diffusion models and allows it to synthesize high-quality novel view images with realistic and complex patterns (Section 3.1).

Moreover, we propose to utilize the pretrained multi-view video diffusion model as a geometrical reference network, extracting high-fidelity geometrical priors to guide the sparse generation process (Section 3.2). The extracted priors are integrated into the sparse multi-view generator via *residual temporal adapter*, which could be seamlessly plugged into the original network architecture without altering the pretrained parameters (Section 3.3). The proposed sparse multi-view generator together with geometrical reference network formulates the overall framework of our model (Section 3.4).

3.1. Sparse Multi-View Generation

Existing methods such as SV3d [33], which are finetuned directly from pretrained video diffusion models, suffer from high computational costs incurred by the large number of viewpoints and fail to produce highly 3D consistent results due to the limited receptive field of the temporal attention. To address these limitations, we propose a novel approach that improves both 3D consistency and generation efficiency by condensing the video diffusion process into a sparser yet more precise generation pipeline. Specifically, we make several key modifications to the original SV3d architecture to adapt it for sparse multi-view generation: i) **Dense 3D Attention**: We extend the original 2D spatial attention layers into dense 3D attention by concatenating keys and values across multiple views, thereby enforcing stronger multi-view consistency. ii) **Ray-based Camera Embedding:** We replace the original 1D camera embeddings with plücker ray embeddings [32], which are concatenated with the input images to enhance the camera control ability. iii) **Camera-aware Frame Embedding**: SV3d originally uses fixed frame embeddings in its temporal attention layers, which limits its flexibility for handling arbitrary camera trajectories and varying numbers of views. Inspired by Zero123 [20], we replace this with a camera-aware frame embedding conditioned on the target camera poses. This modification reduces temporal ambiguity across different views and enables our sparse-view generation network to synthesize novel views under arbitrary camera placements and view numbers.

We initialize the modified multi-view generation model with the parameters of the pretrained video diffusion model and finetune it on a dataset rendered from the ground-truth 3D object dataset Objaverse [9]. During finetuning, we follow SV3d and adopt the EDM [14] framework with a simplified diffusion loss for supervision. However, we observe that directly finetuning the model using a large noise distribution similar to that of SV3d [33] (with $P_{\text{mean}} = 1.2$, $P_{\text{std}} = 1.6$) leads to model collapse and degraded generation quality. To address this problem, we propose a distribution shift strategy by first warming up the training with a smaller noise distribution and then gradually shift to a larger one. Specifically, we set the initial noise distribution as $P^0_{\text{mean}} = 0.6$, $P^0_{\text{std}} = 1.2$ and the end distribution as $P^1_{\text{mean}} = 1.2$, $P^1_{\text{std}} = 1.6$. Given a training step S, the noise distribution P^S is computed as:

$$P^S = \begin{cases} P^0 & S < S^0 \\ \frac{(P^1 - P^0) * (S - S^0)}{(S^1 - S^0)} + P^0 & S^0 \leq S < S^1 \\ P^1 & S^1 \leq S \end{cases} \quad (1)$$

, where S^0, S^1 denote the starting step and end step for the progressive shifting. For clarity, we omit subscripts of P in the equation. Benefit from these carefully designed modifications, our sparse-view generation model retains the generative priors from the pretrained video diffusion model while achieving more efficient and precise generation, being capable of synthesizing highly consistent and realistic multi-view images.

3.2. Dense Geometrical Prior Extraction

While the finetuned sparse multi-view generation model shows relatively good results, we observe that the reduction in view count may degrade the pretrained knowledge, as it disrupts the temporal continuity assumption established during pretraining, leading to unstable generation results for some in-the-wild cases. To address this challenge, we propose a novel strategy that reuses the fine-grained features from the pretrained model to provide geometrical guidance and further enhance the generation stability. We begin by conducting an in-depth analysis of the generation process in pretrained multi-view video diffusion models, and identify key insights for extracting geometrical priors. Features output by the temporal attention layers in the decoder blocks of the video diffusion U-Net provide rich geometrical priors, even at early denoising stages, and remain consistent throughout the denoising process.

Empirical evidence supporting this observation is provided in Figure 3. We visualize the feature maps from different layers of the U-Net in the multi-view video diffusion model. As shown in the right column of Figure 3, the feature maps from the temporal attention layer in the decoder remain highly consistent throughout the denoising process and capture rich geometrical priors even at the early stage of denoising. This insight inspires us to directly utilize the pretrained multi-view video diffusion model as a geometrical reference network to efficiently distill dense geometrical priors. Since the feature maps remain consistent during denoising, we can extract these geometrical priors quickly with just a few denoising steps. In our experiments, we adopt a subsampled eight-step denoise schedule to efficiently capture high-fidelity geometrical information from

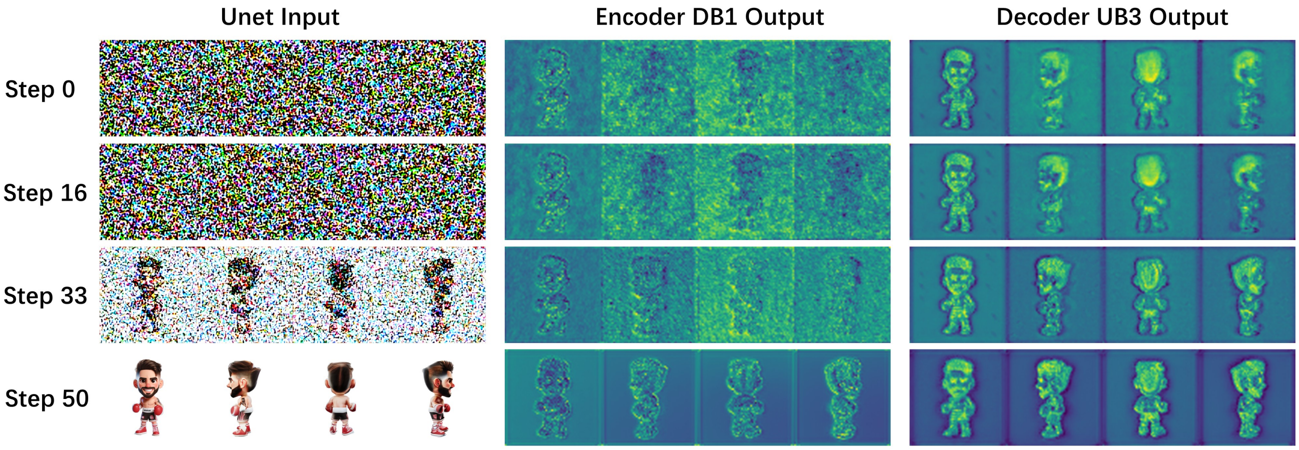

Figure 3. Visualization of feature maps at various stages of the video diffusion model's generation process. From left to right, we show the input to the diffusion model U-Net, the feature maps from the first downsampling block in the encoder, and the feature maps of the third upsampling block in the decoder. As seen in the right column, the feature maps from the temporal attention layer in the decoder remain highly consistent throughout the denoising process. Notably, they capture rich geometrical priors, even at the early stage of denoising.

the pretrained model. The extracted feature maps largely preserve geometrical information and serve as the generative priors to enhance the subsequent generation process.

3.3. Prior Feature Integration

To integrate the extracted geometrical priors into the subsequent sparse multi-view generator, we introduce a lightweight yet effective module, termed residual temporal adapter which could be seamlessly plugged into the original network structure to guide the overall generation process.

The proposed residual temporal adapter consists of two components: a residual temporal attention module and a mask prediction module. Denote the feature map in the sparse multi-view generator as I. We begin by reshaping I by merging the spatial dimensions into the batch axis. The reshaped feature map $I_t \in \mathbb{R}^{(b\times h\times w)\times f\times c}$ (where f denotes the number of synthesized views), along with the extracted geometrical prior features $P \in \mathbb{R}^{(b\times h\times w)\times f_p\times c}$ (where f_p is the frame number generated by the pretrained video diffusion model), are then passed into the residual temporal attention module. The module calculates temporal residuals for each synthesized novel view image, formulated as:

$$\begin{aligned} I_t^{new} &= \text{Softmax}\left(\frac{Q_t K_P^\top}{\sqrt{d}}\right) V_P + I_t, \\ Q_t &= I_t W_q, K_P = P W_k, V_P = P W_v, \end{aligned} \tag{2}$$

where W_* denotes the weights of linear projection layers. Notably, the residual temporal attention incurs low computational and memory overhead, as it operates across views but independently at each spatial location.

In our experiments, we assume the geometrical priors could be extracted from any denoising steps of the pretrained multi-view video diffusion model to enhance the generation of the subsequent sparse-view generator. To modulate the influence of priors extracted at different denoising steps, we propose an adaptive control module to adjust the control strength. This module is implemented with two MLP layers. Let n denote the denoising timestep in the pretrained video diffusion model, and t denote the denoising timestep in the sparse-view synthesis network. The residual attention formulation is then extended as:

$$I_t^{new} = M(n,t) \times \text{Softmax}\left(\frac{Q_t K_P^\top}{\sqrt{d}}\right) V_P + I_t, \tag{3}$$

where $M(n,t)$ is the adaptive control mask predicted by the control module based on both n and t.

During training, only the parameters of the temporal residual adapter are trained while other modules in the sparse multi-view synthesis model are frozen. This paradigm is efficient and preserves the ability and accuracy of the original sparse view synthesis networks. With the proposed temporal residual adapter, geometrical priors from the video diffusion model are effectively captured and integrated into the sparse multi-view generation process, enhancing geometrical stability and generalization ability, and leading to improved generation quality.

3.4. Overall Framework.

The proposed sparse multi-view generator together with the geometrical reference network formulates the overall framework of our model, as illustrated in Figure 2. This framework condenses the multi-view video diffusion process into the generation of high-fidelity sparse multi-view

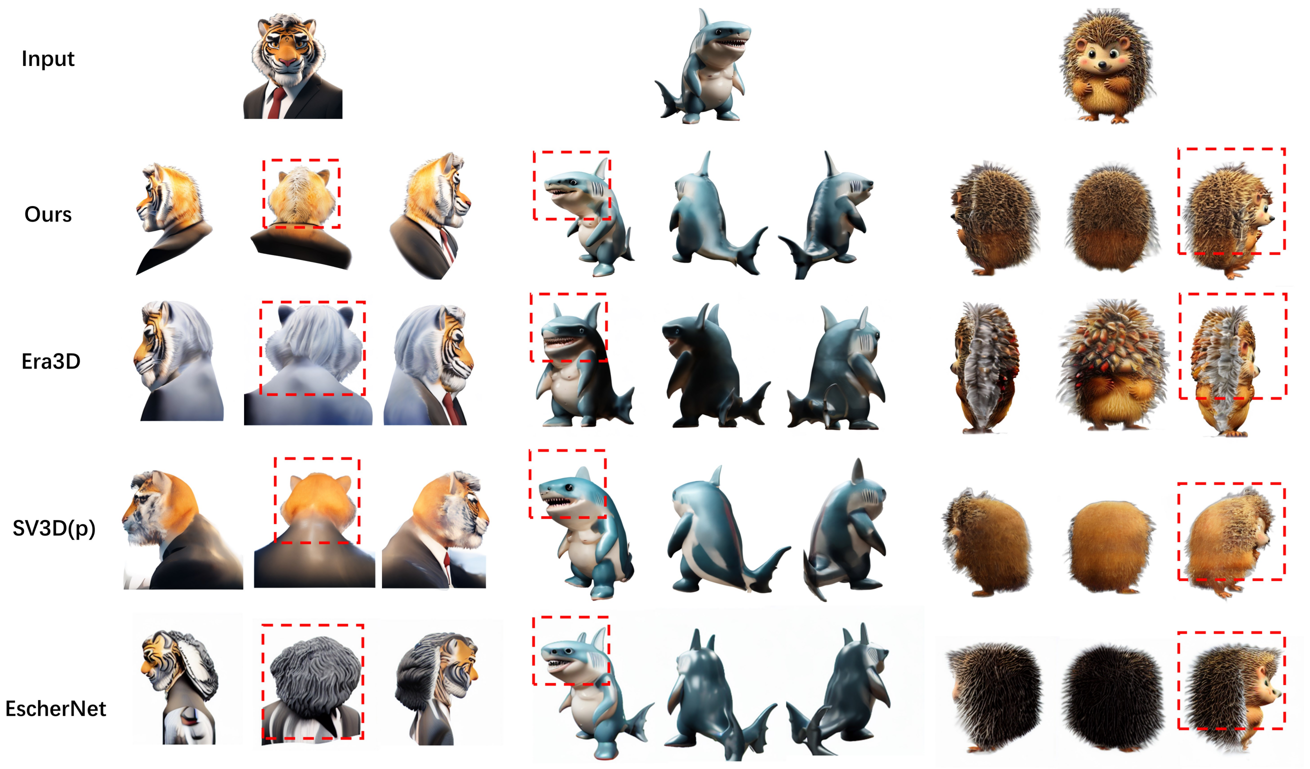

Figure 4. Qualitative comparisons of generated novel views between our models and state-of-the-art novel view synthesis methods. Leveraging strong generative priors inherited from the pretrained video diffusion model, our method synthesizes high-quality novel views with enhanced 3D consistency and realistic visual details.

images, achieving improved 3D consistency and computational efficiency, while largely preserving the generative priors from the pretrained video diffusion model. Leveraging this framework, we are able to synthesize highly consistent sparse multi-view images that can be directly lifted into 3D space using sparse-view reconstruction methods. In our experiments, we adopt GRM [43] as the reconstruction model to lift the generated views into complete 3D objects

4. Experiments

4.1. Implementation

We train our model on the open-source multi-view dataset G-Objaverse [29]. For the base multi-view video diffusion model, we adopt a reproduced version of SV3d [33], which incorporates the plücker ray embedding [32] for camera control. The overall training process for our proposed model consists of two stages. In the first stage, we finetune the multi-view video diffusion model into a sparser yet more precise generator for 30k steps with a batch size of 128. In the second stage, we train the proposed residual temporal adapter, which converges quickly within 10k steps using a batch size of 64. During training, we randomly sample four view as the training objective. We use the AdamW optimizer and FP16 precision for efficient gradient descent, without weight decay. The learning rate for all experiments is set to 1e-5.

4.2. Qualitative Comparison

We provide qualitative comparisons between our proposed method with state-of-the-art multi-view synthesis methods, including EscherNet [15], SV3d(p) [33] and Era3d [16], as shown in Figure 4. For each method, we compute the metrics on three generated novel views with fixed azimuth intervals: 90, 180 and 270 degrees. Due to the limited receptive fields of temporal attention layers, SV3d(p) [33] fails to maintain 3D consistency under large camera movements, resulting in over-smoothed outputs, as illustrated in Figure 4. In contrast, methods such as Era3D [16] and EscherNet [15] suffer from the limited generalization capabilities of 2D diffusion models, resulting in unrealistic outputs when applied to out-of-domain inputs—such as collapsed geometry and incorrect colors in the generated back views.

In contrast, our proposed model effectively preserves the generative priors from the pretrained video diffusion model while enhancing 3D consistency via dense 3D attention. This enables the synthesis of high-quality novel view images with accurate geometry and realistic textures.

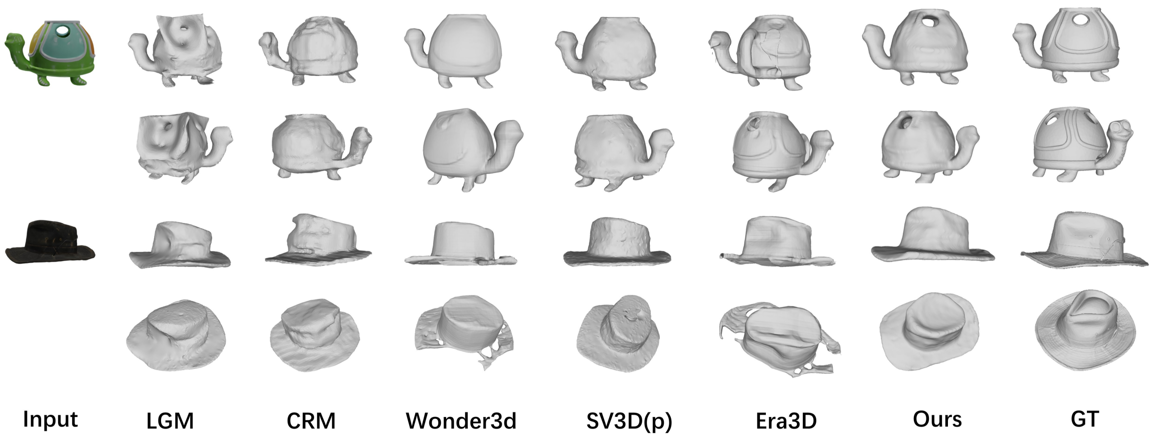

Figure 5. Qualitative comparisons of the generated meshes with existing image-to-3d methods.our model demonstrates strong ability to generate 3D consistent novel view images, which can be reconstructed into high-quality meshes with correct geometric structures and are faithful to the input images.

Dataset	GSO [10]				ABO [8]				OmniObject3D [40]			
Method	PSNR↑	SSIM↑	LPIPS↓	CLIP(S)↑	PSNR↑	SSIM↑	LPIPS↓	CLIP(S)↑	PSNR↑	SSIM↑	LPIPS↓	CLIP(S)↑
Zero123 [20]	17.44	0.8682	0.189	0.803	18.54	0.8791	0.189	0.811	17.28	0.8610	0.213	0.779
Syncdreamer [21]	17.56	0.8674	0.183	0.822	18.69	0.8779	0.180	0.825	17.22	0.8677	0.209	0.786
EscherNet [15]	17.81	0.8712	0.176	0.827	18.94	0.8813	0.171	0.833	18.17	0.8633	0.202	0.791
SV3D(p) [33]	19.88	0.8991	0.119	0.859	20.26	0.9043	0.117	0.864	19.19	0.9021	0.122	0.850
Era3d [16]	20.39	0.9078	0.115	0.869	21.14	0.9131	0.111	0.874	19.97	0.9076	0.119	0.858
Ours(w/o VP)	19.90	0.8993	0.114	0.866	20.54	0.9032	0.113	0.871	19.67	0.8990	0.120	0.854
Ours(w/o DS)	20.23	0.9067	0.111	0.867	20.87	0.9075	0.108	0.877	19.99	0.9025	0.119	0.857
Ours(w/o GR)	20.49	0.9086	0.108	0.874	**21.31**	**0.9143**	0.104	0.880	20.10	0.9090	0.115	0.861
Ours	**20.58**	**0.9103**	**0.106**	**0.879**	21.29	0.9122	**0.101**	**0.887**	**20.20**	**0.9101**	**0.112**	**0.869**

Table 1. Quantitative evaluation between our proposed method with existing novel view synthesis methods. VP denotes Video Pretraining, DS denotes the proposed distribution shift strategy and GR denotes Geometrical Reference Network.

Method	CD↓	IoU↑
Shape-E [13]	0.0651	0.210
One-2-3-45 [19]	0.0516	0.359
Syncdreamer [21]	0.0529	0.361
EscherNet [15]	0.0513	0.382
LGM [32]	0.0425	0.451
CRM [37]	0.0411	0.465
Wonder3d [23]	0.0382	0.468
SV3D(p) [33]	0.0375	0.463
Era3d [16]	0.0369	0.472
Ours	**0.0362**	**0.479**

Table 2. Quantitative comparison between our proposed method with exsting methods on 3D reconstruction.

Furthermore, we provide qualitative comparisons of the final reconstructed meshes. As shown in Figure 5, our model demonstrates strong performance in generating 3D-consistent novel views, which can be reliably reconstructed into high-quality meshes with correct geometric structures that faithfully reflect the input appearance.

4.3. Quantitative Comparisons

We conduct quantitative evaluations on multiple datasets including Google Scanned Objects dataset(GSO) [10], Amazon Berkeley Objects (ABO) Dataset [8] and OmniObject3D [40]. Specifically, for each dataset, we randomly select 200 objects. We use several metrics to assess the quality and multi-view consistency of the generated images at both the pixel and semantic levels. These metrics include Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and CLIP similarity score (CLIP-S). For 3D reconstruction, we select 50 objects from GSO [10] as groundtruth and then evaluate using Chamfer Distances (CD) and Volume Intersection-over-Union (IoU) between the ground-truth shapes and the reconstructed shapes.

Table 1 presents a quantitative evaluation between our proposed model and existing methods on the task of novel view synthesis. Leveraging strong generative priors from the pretrained multi-view video diffusion model, our method outperforms prior approaches in all metrics, demonstrating the effectiveness of our designs. In Table 2, we

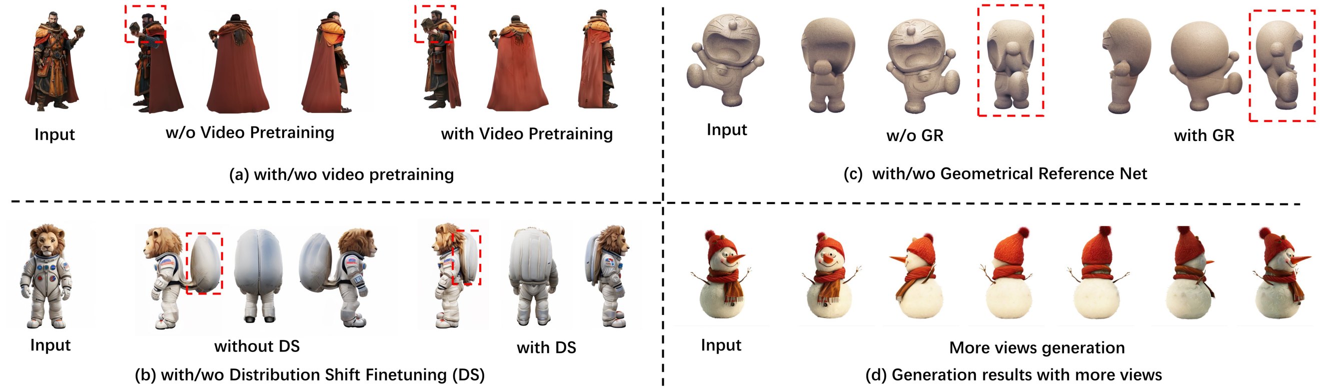
Input
w/o Video Pretraining
with Video Pretraining
(a) with/wo video pretraining
Input
w/o GR
with GR
(c) with/wo Geometrical Reference Net
Input
without DS
with DS
(b) with/wo Distribution Shift Finetuning (DS)
Input
More views generation
(d) Generation results with more views

Figure 6. Ablation studies. We evaluate the effectiveness of video pretraining, distribution shift strategy, geometrical reference net and more views generation.

provide a quantitative comparisons on the metrics of 3D reconstruction. Our model outperforms other baselines, indicating that the generated multi-view images can be reliably lifted into high-quality 3D assets.

4.4. Ablation Studies

Sparse Multi-View Generation with Video Priors. We evaluate the effectiveness of video pretraining and the proposed distribution shift strategy. In the no-video-pretraining scenario, we remove the temporal attention layers, causing the reduced multi-view generator to degrade into a conventional 2D diffusion model. In the without distribution shift scenario, we finetune the model using the same noise distribution as SV3d [33], where $P_{mean} = 1.2$, $P_{std} = 1.6$.

As shown in Figure 6 (a), (b) and Table 1, the model incorporating both video pretraining and the distribution shift strategy achieves the best performance, generating more realistic details and complex patterns. These results highlight the capacity of video diffusion models to provide strong generative priors, significantly improving the realism and consistency of novel view synthesis.

Geometrical Reference Network. As shown in Figure 6 (c) and Table 1, we evaluate the effectiveness of the proposed geometrical reference network. Without this component, the condensed model exhibits geometrical instability and may fail to generate reliable structures, particularly on out-of-domain inputs. In contrast, the geometrical reference network provides rich geometrical priors from the pretrained video diffusion models, enabling the sparse-view generator to synthesize high-fidelity geometry with strong generalization capability.

Number of views. During training, we randomly sample four views as the training objective. At inference time, our model can be directly extended to generate additional views while maintaining strong 3D consistency. In Figure 6 (d), we present results of generating six novel views conditioned on a single input image, demonstrating the model's ability to generalize to a different number of target viewpoints.

5. Limitations

While our framework achieves promising results, its performance remains dependent on the quality of the pretrained video diffusion model. Employing stronger pretrained backbones is expected to further improve generation quality. In addition, the model still struggles with intricate structures, particularly in thin or fine-grained objects. Incorporating stronger 3D reasoning capabilities into the framework could be a promising direction to address this limitation. Finally, our current work focuses on object-centric novel view synthesis. In future work, we plan to extend our framework to more complex settings such as scene-level novel view synthesis.

6. Conclusion

We propose a novel framework that reduces multi-view video diffusion into a sparse-view generation process for high-fidelity novel view synthesis. Specifically, we first condense the video diffusion model to synthesize highly consistent sparse multi-view images and then extracting dense geometrical priors from the pretrained video diffusion to enhance the generation stability, which formulates a novel framework that is capable of synthesizing highly consistent sparse multi-view images. Extensive experiments on multiple datasets demonstrate that our model generates highly consistent novel views with superior 3D consistency and significantly improved generation efficiency.

Acknowledgement

This research is supported by the MoE AcRF Tier 2 grant (MOE-T2EP20223-0001) and the MoE AcRF Tier 1 grant (RG14/22).

References

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

[2] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2

[4] Chi Chen, Ang Jin, Zhiye Wang, Yongwei Zheng, Bisheng Yang, Jian Zhou, Yuhang Xu, and Zhigang Tu. Sgsr-net: structure semantics guided lidar super-resolution network for indoor lidar slam. *IEEE transactions on multimedia*, 26: 1842–1854, 2023. 2

[5] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning 3d shape latent for point cloud completion. *IEEE Transactions on Multimedia*, 2024. 2

[6] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 1, 3

[7] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 2

[8] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 7

[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*, pages 13142–13153, 2023. 4, 1

[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. 7

[11] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 2

[12] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 2

[13] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2, 7

[14] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 4

[15] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 6, 7

[16] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 1, 3, 6, 7

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

[18] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[19] 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. 7

[20] 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. 3, 4, 7

[21] 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. 7

[22] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023. 2

[23] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1, 3, 7

[24] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2

[25] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 2

[26] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

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

[28] 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. 3

[29] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9925, 2024. 6

[30] 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. 3

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

[32] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3, 4, 6, 7

[33] 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. *arXiv preprint arXiv:2403.12008*, 2024. 1, 2, 3, 4, 6, 7, 8, 5

[34] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1

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

[36] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[37] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pages 57–74. Springer, 2025. 7

[38] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 3

[39] 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. 3

[40] 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. 7

[41] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 3

[42] 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. 3

[43] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 1, 3, 6

[44] Fan Yang, Jianfeng Zhang, Yichun Shi, Bowen Chen, Chenxu Zhang, Huichao Zhang, Xiaofeng Yang, Jiashi Feng, and Guosheng Lin. Magic-boost: Boost 3d generation with mutli-view conditioned diffusion. *arXiv preprint arXiv:2404.06429*, 2024. 3

[45] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Chong-Wah Ngo, and Tao Mei. Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. *arXiv preprint arXiv:2409.07452*, 2024. 1

[46] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5826–5835, 2021. 2

[47] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*, 2024. 1, 3