This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FVGen: Accelerating Novel-View Synthesis with Adversarial Video Diffusion Distillation

Wenbin Teng[1,2], Gonglin Chen[1,2], Haiwei Chen[1,2], Yajie Zhao[1,2*]
[1]Institute for Creative Technologies [2]University of Southern California
{wenbinte, gonglinc}@usc.edu, {chenh, zhao}@ict.usc.edu,

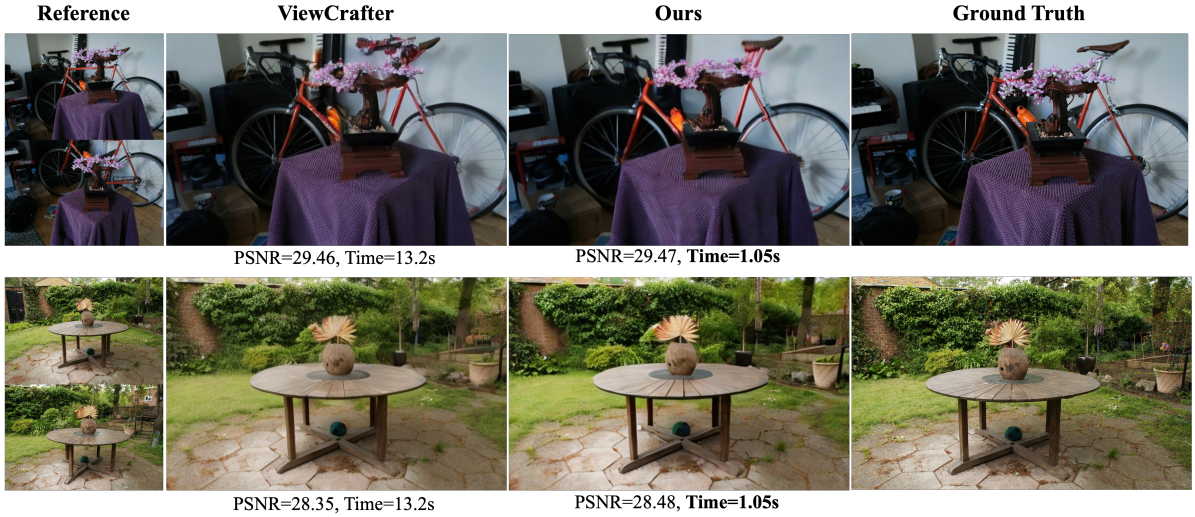

Figure 1. We present **FVGen**, a method of fast novel-view synthesis for 3D scene reconstruction with sparse inputs. Recent method [52] unleash the generation capabilities of video diffusion models conditioned on a prior 3D cue for dense view creation, but always suffers from a long sampling time. We propose a novel framework that can significantly reduce the sampling time and maintain the generation quality.

Abstract

Recent progress in 3D reconstruction has enabled realistic 3D models from dense image captures, yet challenges persist with sparse views, often leading to artifacts in unseen areas. Recent works leverage Video Diffusion Models (VDMs) to generate dense observations, filling the gaps when only sparse views are available for 3D reconstruction tasks. A significant limitation of these methods is their slow sampling speed when using VDMs. In this paper, we present FVGen, a novel framework that addresses this challenge by enabling fast novel view synthesis using VDMs in as few as four sampling steps. We propose a novel video diffusion model distillation method that distills a multi-step denoising teacher model into a few-step denoising student model using Generative Adversarial Networks (GANs) and softened reverse KL-divergence minimization. Extensive experiments on real-world datasets show that, compared to previous works, our framework generates the same number of novel views with similar (or even better) visual quality while reducing sampling time by more than 90%. FVGen significantly improves time efficiency for downstream reconstruction tasks, particularly when working with sparse input views (more than 2) where pre-trained VDMs need to be run multiple times to achieve better spatial coverage. Project Page: https://wbteng9526.github.io/fvgen/

1. Introduction

Reconstructing 3D scenes from 2D images remains a pivotal research area at the intersection of computer vision and graphics. This field has extensive social impact, with broad applications in autonomous navigation, augmented reality, and more. With the recent development of neural rendering techniques such as NeRF [29] and 3DGS [16], creating high-quality 3D assets from 2D images has become more

*Corresponding author.

accessible. However, the success of 3D reconstruction depends on dense observations, which are often difficult to obtain in real-world scenarios. This limitation makes it challenging to apply these methods in many practical applications.

Introducing generative models into 3D reconstruction can fill gaps in missing views when only sparse input views are available. However, novel view synthesis from sparse views, especially in the settings with less than 2 views, is inherently an ill-posed problem. A bulk of previous works [7, 8, 30, 43, 52] have found that priors learned from large image and video diffusion models [10, 11] are rich enough to inpaint detailed and reasonable information in the many under-observed areas in the reconstructed 3D representations. We are primarily interested in the use of video diffusion models (VDM) to perform novel view synthesis. What bridges VDM to novel view synthesis is its ability to synthesize continuous views following a camera trajectories between the observed views. The spatio-temporal consistency in these smooth video trajectories, as demonstrated in several works [6, 7, 25, 44, 52], effectively addresses the challenging "Janus Problem" that typically limits generation quality in other multi-view settings. However, VDM is a computationally expensive method because of the iterative sampling process that characterized all denoising diffusion methods. The long generation time makes VDM-based generation methods unsuitable for certain real-world downstream tasks, such as dynamic 3D reconstruction, where new consistent views must be generated continuously. This limitation also affects large-scale 3D reconstruction that requires synthesizing a significant number of novel views.

In this paper, we show that VDM-based novel view synthesis can be significantly accelerated for free. Without any compromise on the generation quality, we speed up diffusion-based synthesis of novel view by a factor of 90% with a 4-step student model trained under a GAN objective [23]. Acceleration of diffusion models has been actively studied in the past, but previous techniques may not be well suited to our task due to two reasons: First, we need a method specifically for accelerating video diffusion models, rather than image diffusion models. Second, the method must generalize effectively when trained on multi-view video datasets, which contain significantly less data than general video datasets. To accelerate a diffusion model, recent methods [22, 28, 34, 46, 49] propose minimizing the distributional difference between a few-step student generation model and the original multi-step teacher model. However, these methods are restricted to images with specific domains. The most related work to us is [50], which generates more video frames with faster speed by optimizing a distribution matching distillation (DMD) loss. However, this method requires an effective initialization of the student model. This is completed by generating a noise-latent pair through an ODE solver [36] and training the student model by optimizing a regression loss [49, 50]. The generation process usually takes a very long time and the student's ability is limited by the teacher's generation quality. More importantly, we have observed that DMD optimizations tend to be unstable and have the tendency towards mode collapse, when trained under multi-view video data. This may be due to the inherent concept of DMD loss optimization being the reverse KL-divergence minimization, which tends to be mode-seeking and zeroing out the modes of teacher's that are not relatively dominant.

To solve the previous problems with VDM acceleration, we propose a novel framework, named **FVGen**, that performs novel views generation with a fast video diffusion generation with as few as 4 steps. First of all, different from previous GAN-based distillation methods [34, 51], we find it beneficial to initial the few-step student model by training a GAN objective, where the student model is regarded as a generator of fake samples, and, instead of training a separate discriminator as in previous works, we leverage a pre-trained teacher model as the discriminator. Second, we propose the soften reverse KL-divergence to solve the unstable optimization of DMD loss. The soften reverse KL-divergence maintains similar mode-seeking behavior as reverse KL-divergence but with more robust distillation: it prevents the student from ignoring entire space of teacher's distribution, thereby preserving more of teacher's knowledge, which we find particularly beneficial in the context of limited data.

Through effective and robust distillation, FVGen is capable of generating video sequences with comparable or better visual quality as ViewCrafter [52] but with significantly less time. FVGen is trained with Dl3DV-10K [24] dataset and extensive experiments on real world dataset including MipNeRF360 [1] and Tanks-and-Temples [18] demonstrate that our method achieves overall better performance than other state-of-the-art (SOTA) novel view synthesis methods and diffusion distillation methods.

2. Related Works

2.1. Diffusion-based Novel View Synthesis

NeRF [29] and 3DGS [16] excel at novel view synthesis with dense input views. However, reconstructing scenes from sparse views requires additional priors. Diffusion models [10, 32] can generate realistic pseudo input views from sparse original inputs. Several works use Score-Distillation Sampling [21, 30, 39, 42] to create 3D objects from text prompts or single images by distilling latent diffusion model [32] knowledge into 3D representations. While NeRF, 3DGS, and SDS typically require time-consuming per-scene optimization, view-dependent diffusion models [8, 19, 20, 35] offer an alternative. These mod-

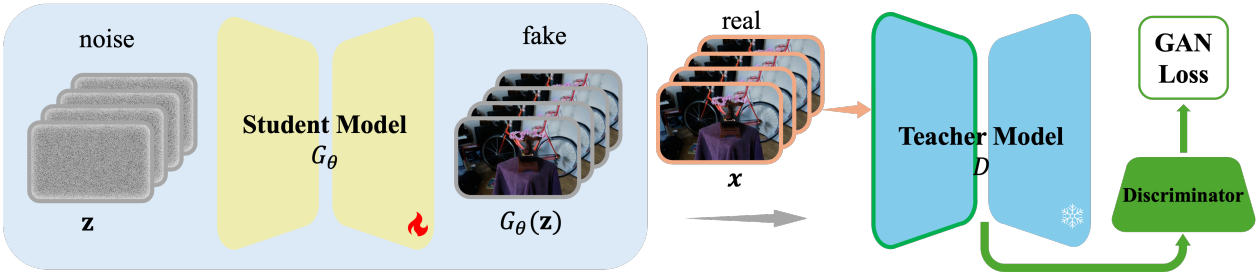

Figure 2. **Overview of student initialization.** We initialize our student model by training a GAN objective. The student model G_θ, initialized with the weights of teacher model, uses few-step denoising to generate fake videos to fool the discriminator D. The fixed teacher model, together with a trainable discriminator, are optimized to differentiate between fake and real videos.

els directly generate 3D-consistent novel views conditioned on input images but directly lifting 2D images to 3D input therefore face challenges when scaling to larger scenes.

An alternative approach leverages video diffusion models [6, 7, 25, 38, 52] for multi-view image generation. These methods fine-tune pre-trained latent video diffusion models (LVDM) [2, 4, 12, 44, 48] using multi-view images with both 2D and 3D guidance. The 2D guidance typically involves semantic features like CLIP [14] embeddings, while 3D guidance comes from sparse input cues. For instance, [52] uses DUSt3R [40] to build an initial point cloud and guides generation by combining point cloud renders with camera poses. Similarly, [6] conditions generation on feed-forward 3DGS renders, enabling 360-degree scene generation with longer frame sequences through multi-step refinement. However, these methods suffer from speed issues caused by video diffusion sampling when generating novel views.

2.2. Diffusion Distillation

Diffusion models typically require numerous denoising steps to generate high-quality samples, which makes them computationally intensive. To accelerate generation, distillation methods train a student model that mimics the teacher diffusion model's behavior while using fewer sampling steps. Using fewer sampling steps reduces quality, so existing works incorporate adversarial training [3, 15, 22, 27, 33, 34, 46] to enhance the student model's output. However, simple GAN training cannot fully bridge the gap between student and teacher distributions, causing the student's generated output to remain noticeably different from the original diffusion model. To address this problem, Distribution Matching Distillation (DMD) [49, 51] optimizes the reverse KL-divergence. For tractable optimization, [49, 51] leverages the gradient of DMD loss, which represents the difference between student and teacher score functions. For fast video generation, [50] extends DMD to video diffusion models like CogVideoX [48]. While these methods work well for general image and video generation tasks, they suffer from mode collapse and training instability when applied to view synthesis on relatively small datasets like multi-view video datasets.

3. Preliminaries

3.1. Video Diffusion Model

A video diffusion model is a generative model designed to synthesize realistic video sequences by learning the complex temporal and spatial patterns within video data. Based on the diffusion process, these models iteratively apply noise to video frames and then learn to reverse this process to generate new sequences. The forward diffusion process progressively adds Gaussian noise to the input clean sample $\boldsymbol{x}_0 \sim p(\boldsymbol{x})$ over T timesteps, leading to highly noisy data $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$. This process is formulated as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\boldsymbol{I}), \quad (1)$$

where α_t and $\boldsymbol{x}_t$ are the noise strength and noisy data at timestep t. The model then learns a denoising function, often parameterized by a neural network $\epsilon_\theta(\boldsymbol{x}_t, t)$, to reverse this process and generate realistic video frames, where the optimization function is defined by the MSE loss:

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(0,I),\boldsymbol{x}\sim p,c}\left\|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t, c)\right\|^2, \quad (2)$$

where c is the condition embedding of diffusion model which is usually represented as a text or image prompt. Instead of generating sequences at full resolution, Latent Video Diffusion Model (LVDM) first encodes video data $\boldsymbol{x} \in \mathbb{R}^{F\times C\times H\times W}$ into latent space using a pre-trained VAE [17] encoder: $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$, where $\boldsymbol{z} \in \mathbb{R}^{F\times C\times h\times w}$. The compression will mitigate computational complexity while maintaining video generation quality.

3.2. Distribution Matching Distillation

Distribution matching distillation adopts the idea of variational score distillation [42] that is optimized to match the distribution of student and teacher model by minimizing the

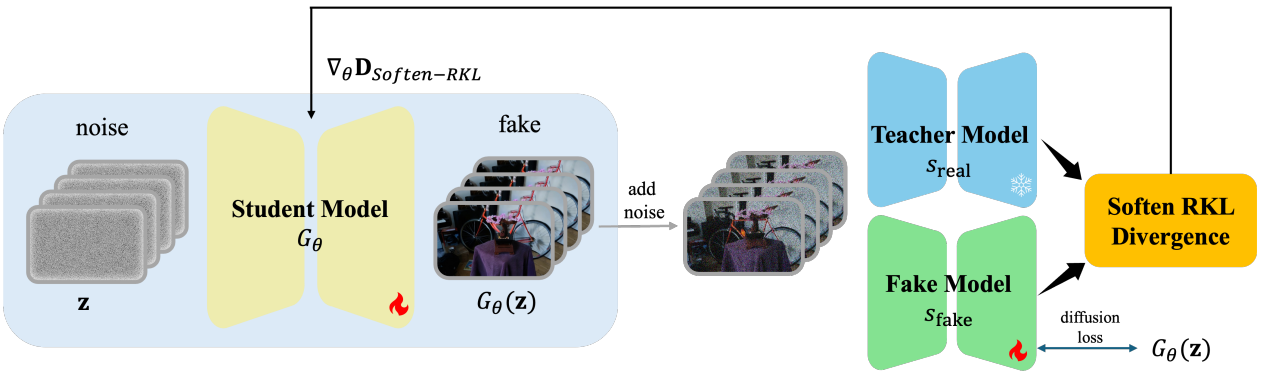

Figure 3. **Overview of Distribution Matching Distillation (DMD).** The gradient is the difference between the teacher score and fake score w.r.t. the student model output. We apply soften reverse KL-divergence for stable training and avoiding mode-collapse.

reverse KL-divergence. Given a student model G parameterized by θ, the gradient of reverse KL-divergence is

$$\begin{aligned}\nabla \mathcal{L}_{\text{DMD}} &= \mathbb{E}_t \left(\nabla_\theta \mathbf{D}_{\text{Reverse-KL}}(p_{\text{fake},t} \parallel p_{\text{real},t})\right) \\ &= -\mathbb{E}_t \Bigg(\Big[s_{\text{real}}(F(G_\theta(\boldsymbol{z}), t), t) \\ &\quad - s_{\text{fake}}(F(G_\theta(\boldsymbol{z}), t), t)) \frac{dG_\theta(\boldsymbol{z})}{d\theta} \Big] \Bigg), \end{aligned} \tag{3}$$

where p_{real} and p_{fake} are the real and fake distribution. $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$ is the initial random Gaussian noise. F is the diffusion forward function determined by Eq. (1). s_{real} and s_{fake} are the real and fake score function and defined as follows according to [37]:

$$s(\boldsymbol{x}_t, t) = \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = -\frac{\boldsymbol{x}_t - \alpha_t \mu(\boldsymbol{x}_t, t)}{\sigma_t^2}, \tag{4}$$

where p_t is the distribution of the noisy sample $\boldsymbol{x}$. μ is the clean sample prediction of the diffusion model. While optimizing student model G_θ with gradient descent formulated in Eq. (3), the gradients of μ_{real} are stopped and μ_{fake} is dynamically optimized with the output of G_θ with regular diffusion loss similar to Eq. (2)

4. Method

We propose Fast Video Generation (**FVGen**) that distills the generation capability of a multi-step teacher 3D-VDM into a few-step student 3D-VDM. The teacher 3D-VDM used in our experiments is the novel-view synthesis model ViewCrafter [52]. From a set of observed images $\boldsymbol{I}^{\text{obs}}$ and observed cameras $\boldsymbol{\pi}^{\text{obs}}$, ViewCrafter learns a conditional distribution $\boldsymbol{I}^{\text{tgt}} \sim p(\boldsymbol{I}^{\text{tgt}} | \boldsymbol{I}^{\text{obs}}, \boldsymbol{\pi}^{\text{obs}}, \boldsymbol{\pi}^{\text{tgt}})$, where $\boldsymbol{I}^{\text{tgt}} = \{\boldsymbol{I}'_i\}_{i=1}^{L}$ is a sequence of L images corresponding to the target views $\boldsymbol{\pi}^{\text{tgt}}$. The student model FVGen is trained to sample from the same conditional distribution with a much faster inference speed. The architectures of FVGen are shown in Figure 2 and 3. For notation simplicity purpose, we ignore including $\boldsymbol{I}^{\text{obs}}, \boldsymbol{\pi}^{\text{obs}}, \boldsymbol{\pi}^{\text{tgt}}$ in the future formula as they are all the conditional inputs for both teacher and student diffusion model.

We first initialize the parameters of student model by training with a generative adversarial network (GAN) objective, leveraging the teacher model as discriminator (Section 4.1). After that, we continue to train the student model by optimizing DMD loss and diffusion loss to minimize the soften reverse KL-divergence between the distribution of student and teacher model (Section 4.2).

4.1. Student Initialization with GAN Training

Directly training student model with DMD (Eq. 3) will cause training collapse as it is difficult to naively distill multi-step denoising into few-step. [49, 50] proposed to create noise-latent pairs by an ODE solver [36] and minimize a regression loss with student model for initialization. However, we found that it is very time consuming to generate large amount of video data pairs and the student model will also be restricted by the teacher's limitations. Inspired by [23, 51], we propose to initialize the student model by training a GAN objective with the real images. Specifically, the student model (G_θ) generates fake video samples that fool the teacher model who serves as a discriminator (denoted as D) by minimizing the generator loss $\mathcal{L}_G$. The discriminator D classifies real samples from generated fake samples by maximizing $-\mathcal{L}_D$. The min-max game adversarial optimization is formulated as:

$$\begin{aligned}\mathcal{L}_D &= \mathbb{E}_{\substack{\boldsymbol{x}\sim p_{\text{real}} \\ t\sim\mathcal{U}[0,T]}} \left[\log f\left(D(F(\boldsymbol{x}, t))\right)\right] \\ &\quad - \mathbb{E}_{\substack{\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}) \\ t\sim\mathcal{U}[0,T]}} \left[\log f\left(D(F(G_\theta(\boldsymbol{z}), t))\right)\right], \\ \mathcal{L}_G &= \mathbb{E}_{\substack{\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}) \\ t\sim\mathcal{U}[0,T]}} \left[\log f\left(D(F(G_\theta(\boldsymbol{z}), t))\right)\right], \end{aligned} \tag{5}$$

where t is the time-step sampling and F is the forward noise-injection model similar to Eq. (3). Discriminator D retrieves the output feature map of the middle block of diffusion model UNet and passes it into the classifier f. To

fit with video data, f is composed of a 3D Convolutional Neural Network that compresses the feature map into classification logits. Here, we use the original sample $\boldsymbol{x}$ as the input to the discriminator instead of the generated sample $\boldsymbol{x}$ from teacher model such that the student model would not be restricted by the generation capability of teacher model.For the purpose of training stability, recent methods [6, 52] predict the velocity field $\boldsymbol{v}_\theta$, and we convert it into the sample prediction with $\boldsymbol{x}_0 = \sqrt{\alpha_t}\boldsymbol{x}_t - \sqrt{1-\alpha_t}\boldsymbol{v}_\theta$. Although the generated video samples are still very blurry, the GAN training scheme provides an effective initialization of the student model, which paves the way for more effective training with distribution matching loss.

4.2. Distribution Matching with Soften Reverse KL-Divergence Minimization

As discussed in Section 3.2, the core concept of distribution matching distillation (DMD) [49, 51] is to minimize the reverse KL-divergence between student and teacher distributions. While reverse KL is popular for its mode-seeking focus, it has notable limitations in the context of generative modeling, its tendency toward mode collapse. Since reverse KL rewards the student for zeroing out probability in any region where the teacher's density is low, the student can effectively ignore parts where the teacher's distribution are not dominant. If the teacher's distribution $p_{\text{real}}(\boldsymbol{x})$ has multiple modes of varying height, minimizing Reverse-KL$(p_{\text{fake}} \| p_{\text{real}})$ may lead p_{fake} to concentrate on a subset of those modes and assign negligible mass to others. Therefore, the exclusive nature of reverse KL-divergence makes it prone to mode dropping and a lack of coverage of the full distribution.

To address the shortcomings of pure reverse KL, we propose to leverage a $Soften$ reverse KL-divergence (Soften RKL) [9, 47]. Soften RKL refers to a modified divergence that retains the mode-seeking bias of reverse KL but softens its exclusion other relatively lower-probability regions, thereby mitigating mode collapse. Compared to reverse KL formulated in Eq. (3), the Soften RKL can be formulated as:

$$\begin{aligned}\mathbf{D}_{\text{Soften-RKL}}(p_{\text{fake},t} \| p_{\text{real},t}) &= \mathbf{D}_{KL}\left(\frac{1}{2}p_{\text{real}} + \frac{1}{2}p_{\text{fake}} \middle\| p_{\text{real}}\right), \\ &= p_{\text{fake}}(r+1)\log\left(\frac{1}{2} + \frac{1}{2r}\right),\end{aligned} \tag{6}$$

where $r = r(\boldsymbol{x}, t) := p_{\text{real},t}(\boldsymbol{x}) / p_{\text{fake},t}(\boldsymbol{x})$ is the density ratio. Here $\boldsymbol{x}$ denotes the noise-injection version of the output of student model formulated as $F(G_\theta(\boldsymbol{z}), t)$ with $t \sim \mathcal{U}(0, T)$. Thanks to our optimized GAN objective in Section 4.1, the density ratio could be approximated by the following:

$$r(\boldsymbol{x}, t) = \frac{p_{\text{real},t}(\boldsymbol{x})}{p_{\text{fake},t}(\boldsymbol{x})} \approx \frac{f(D(\boldsymbol{x}, t))}{1 - f(D(\boldsymbol{x}, t))}. \tag{7}$$

In summary, the well-optimized GAN objective not only provides an efficient weight initialization of the student model, but also provides a direct estimate of the density ratio. Here, instead of directly comparing p_{fake} and p_{real} as in $\mathbf{D}_{KL}(p_{\text{fake}} \| p_{\text{real}})$, we compare an even mixture of p_{real} and p_{fake} to p_{real}. Soften RKL maintains the mode-seeking property of reverse KL, while penalizes more on mode dropping. This criterion yields a more robust distillation since it prevents the student from ignoring the entire teacher distribution, thereby preserving more of teacher's knowledge. Through experiments, we find that the soften RKL yields more stable training and more realistic generations. Computing the divergence directly is generally intractable, the gradient with respect to θ is formulated as follows:

$$\begin{aligned}&\nabla \mathcal{L}_{\text{DMD}}^{\texttt{Soften-RKL}} = \\ &-\mathbb{E}_t\left(\frac{1}{r(\boldsymbol{x}, t)}\Big[s_{\text{real}}(\boldsymbol{x}, t) - s_{\text{fake}}(\boldsymbol{x}, t)\Big]\frac{dG_\theta(\boldsymbol{z})}{d\theta}\right),\end{aligned} \tag{8}$$

where $\boldsymbol{x} = F(G_\theta(z), t)$ as previously defined. Apart from the soften RKL divergence optimization, we dynamically update the fake diffusion model μ_{fake} to adjust to change of student distribution in accordance with the original setting of DMD [49]. The loss term is identical to Eq. (2). We summarize the full training procedure in Alg. 1

Algorithm 1 FVGen Training Procedure

Require: Few-step timesteps $\mathcal{T} = \{0, t_1, t_2, \ldots, t_Q\}$, pre-trained teacher model μ_{real}, discriminator classifier D dataset $\mathcal{D}$.

Initialize student model G_θ with μ_{real}.
Initialize fake score function with μ_{fake}.
while training **do**
  Sample a video from dataset $\boldsymbol{x}_0 \sim \mathcal{D}$
  Add noise with timestep $t \sim \mathcal{T}$: $\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
  Predict with student: $\hat{\boldsymbol{x}}_0 = G_\theta(\boldsymbol{x}_t, t)$
  Add noise $\boldsymbol{\epsilon}' \sim \mathcal{N}(0, \boldsymbol{I})$ with timestep $\tau \sim \mathcal{U}(0, T)$ to real and fake sample:
  $\hat{\boldsymbol{x}}_\tau = \sqrt{\alpha_\tau}\hat{\boldsymbol{x}}_0 + \sqrt{1-\alpha_\tau}\boldsymbol{\epsilon}'$
  $\boldsymbol{x}_\tau = \sqrt{\alpha_\tau}\boldsymbol{x}_0 + \sqrt{1-\alpha_\tau}\boldsymbol{\epsilon}'$
  Update G_θ and D with Eq. (5)
end while
Output trained student G_θ
while training **do**
  Repeat step 4 - 8
  Update G_θ with DMD Loss: Eq. (8)
  Add noise $\boldsymbol{\epsilon}'' \sim \mathcal{N}(0, \boldsymbol{I})$ with $t_1 \sim \mathcal{U}(0, T)$ to $\hat{\boldsymbol{x}}_0$:
  $\hat{\boldsymbol{x}}_{t_1} = \sqrt{\alpha_{t_1}}\hat{\boldsymbol{x}}_0 + \sqrt{1-\alpha_{t_1}}\boldsymbol{\epsilon}''$
  Update μ_{fake} with denoising loss: Eq. (2)
end while

Figure 4. **Qualitative Results.** We compare FVGen with SOTA novel view synthesis method on MipNeRF360 [1] and Tanks-and-Temples [18] datasets.

5. Experiments

5.1. Experiment Setup

Implementation Details. All of our student model, teacher model and fake score function is initialized with the parameters of the sparse model of ViewCrafter [52]. We first initialize the student model by training GAN objective. The parameters of teacher model is fixed, and the student model and GAN classifier are trained for 4000 iterations with a two-scale update rule inspired by [51]. Next, we optimize DMD loss by continuously training student model and the fake score function for another 5000 iterations with the same two-scale update rule. The whole pipeline is trained on 8 NVIDIA H100 with batch size 4. The training last about 1 day. We utilize AdamW optimizer [26] and set learning rates of all optimizers to be 5×10^{-5}. The model is trained on images with a resolution of 512×320.

Datasets. FVGen is trained with DL3DV-10K [24], a real-world, scene-level video dataset that includes more than 10K long video clips. Similar to [52], we construct point cloud and ground truth video pairs using DL3DV dataset. We randomly sample short video clips with a random frame stride less than 5. The first frame and last frame are used to construct the prior point cloud with DUSt3R [40]. We use PyTorch3d [31], an efficient renderer to construct the point cloud renders with the camera poses of intermediate frames. Through this process, we create 20,000 training data pairs. We validate our method on two public datasets Tanks-and-Temples [18] and MipNeRF360 [1]

Baseline Methods and Evaluation Metrics. To evaluate our proposed method, we compare our method with several state-of-the-art methods: ViewCrafter [52], MVSplat360 [6]. We also compare our video diffusion distillation method with two state-of-the-art methods: CausVid [50] and DMD2 [51]. We use PSNR [13], SSIM [41], LPIPS [53] to measure reconstruction quality. In addition, we also reported the distribution metric, i.e. Frechet Inception Distance (FID), which compares the distribution of generated views and ground truth views. Apart from the visual quality comparison.

5.2. Novel View Synthesis Comparison

Similar to the evaluation setting of MVSplat360 [6], for each scene of both MipNeRF360 [1] and Tanks-and-Temples [18], we select 5 views that are far from each other but could cover the whole scene. We use our method and baseline method to generate 56 views sampled from the natural camera trajectory (14 views between each 2 input views).

Qualitative Results The qualitative comparisons with SOTA novel view synthesis methods are shown in Figure 4. DepthSplat is a feed-forward generalizable Gaussian splatting method. When there are relatively low overlap between input frames, DepthSplat would exhibit ob-

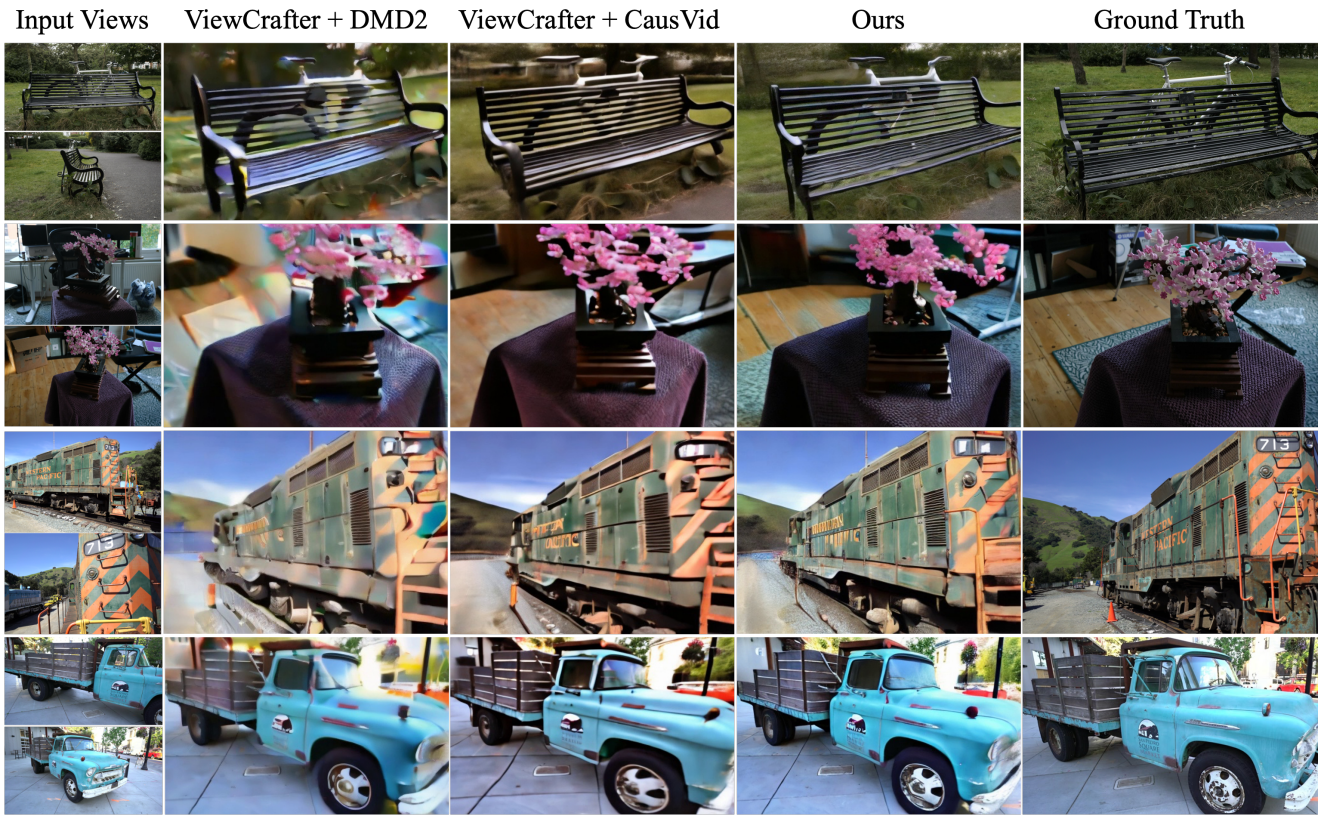

Figure 5. **Qualitative Results.** We compare FVGen with baseline diffusion distillation method on MipNeRF360 [1] and Tanks-and-Temples [18] datasets.

vious artifacts due to inconsistent depth scale and the issue of floating Gaussians. In comparison, MVSplat360 [6] uses VDM to refine the Gaussian splatting renders and remove the floaters. However, the current MVSplat360 only supports training with low resolution inputs and the current results present significant blurriness. Compared to DepthSplat and MVSplat360, ViewCrafter [52] generates photo-realistic novel views with higher resolution. However, ViewCrafter is limited by the generation speed. Given generating 16 total frames, the average inference time is 13.2 seconds. In comparison, our method achieves a 10x speedup, generating the same number of frames in approximately 1 second and achieves similar visual results. The qualitative results prove that our method is able to distill the multi-step sampling into a few-step sampling process and achieve results that are similar to or even better than teacher ViewCrafter model with the student model.

Quantitative Analysis Table 1 presents quantitative comparisons on both datasets of FVGen and other novel view synthesis methods. Except for ViewCrafter [52], our method significantly surpasses other SOTA methods in terms of both perceptual and distribution metrics. Compared with ViewCrafter, our method yields similar quality but with more than 90% faster in generation speed. The quantitative results also highlights the effectiveness of our proposed method.

		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	Time ↓
mip-360	DepthSplat	11.23	0.213	0.715	32.45	4.2
	MVSplat360	12.28	0.285	0.682	25.69	87.2
	ViewCrafter	16.35	0.346	0.433	16.28	66.3
	Ours	16.28	0.352	0.429	17.44	5.1
TNT	DepthSplat	12.43	0.263	0.677	35.88	4.3
	MVSplat360	14.18	0.301	0.532	25.23	87.3
	ViewCrafter	18.69	0.402	0.208	23.94	65.9
	Ours	18.72	0.411	0.210	23.64	5.0

Table 1. **Quantitative Results.** We report four quantitative metrics for image quality comparison on 2 separate datasets: MipNeRF360 [1] (denoted as mip-360 in the table) and Tanks-and-Temples [18] (denoted as TNT in the table). We highlight the best results in red and second-best in yellow.

5.3. Diffusion Distillation Comparison

We also compare FVGen with SOTA diffusion distillation methods. Similar to the 2-view evaluation setting of DepthSplat [45] and MVSplat [5], we select 2 views that are far from each other in both MipNeRF360 [1] and Tanks-and-

Temples [18] dataset. We use our method and baseline method to generate 16 views samples from the natural camera trajectory (14 intermediate views and 2 input views).

Qualitative Results First we perform DMD2 [51] on ViewCrafter for few-step distillation. DMD2 applies end-to-end training of student model, GAN discriminator and fake score function. We find that this training scheme is not stable and present significant variance so that it is hard to converge, leading to blurry results as visually depicted in Figure 5. In addition, we compare our method with a recent work, Caus-Vid [50], on ViewCrafter distillation. Caus-Vid [50] creates a small dataset with ODE solver and initialize the student model by optimizing a regression loss. Caus-Vid also leverages DMD [49] for 4-step video generation. However, the qualitative results indicate that our GAN object training scheme provides better initialization compared to the regression training scheme of ODE pairs.

		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
mip-360	VC+DMD2	9.29	0.184	0.836	35.59
	VC+CausVid	15.77	0.336	0.441	19.27
	Ours	**16.28**	**0.352**	**0.429**	**17.44**
TNT	VC+DMD2	10.27	0.209	0.731	37.67
	VC+CausVid	17.33	0.405	0.232	24.92
	Ours	**18.72**	**0.411**	**0.210**	**23.64**

Table 2. **Quantitative Results.** We report four quantitative metrics for image quality comparison on 2 separate datasets: Mip-NeRF360 [1] (denoted as mip-360 in the table) and Tanks-and-Temples [18] (denoted as TNT in the table). **VC** stands for ViewCrafter [52] for space saving purpose. **VC+X** means applying diffusion acceleration method X on ViewCrafter.

Quantitative Analysis Table 2 presents quantitative comparisons with other diffusion distillation methods, and further proves the superiority of our method. Our method differs with DMD2 [51] in several ways: 1) we apply a 3D discriminator on video data, 2) we perform soften reverse-KL divergence for DMD optimization, 3) our GAN training and DMD training are kept separate for more stable training and more accurate density ratio calculation. Through training DMD2 [51], we found a unstable training process and we recorded the last model before the training collapsed. In comparison, CausVid [50] presents more stable training than DMD2 [51] but still suffer from mode collapse. Our method surpasses the other two baseline methods in both perceptual and distribution metrics.

5.4. Ablation Studies

In this section, we analyze the contributions of each module of our proposed method in Table 3 and Figure 6.

GAN	DMD	Soften RKL	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
✗	✓	✓	8.62	0.154	0.880	40.17
✓	✗	✗	16.23	0.369	0.375	21.48
✓	✓	✗	16.85	**0.385**	0.337	21.05
✓	✓	✓	**17.50**	0.382	**0.320**	**20.54**

Table 3. **Ablation studies.** Quantitative analysis of different components of FVGen.

Figure 6. **Ablation Studies.** Visualization of the contribution of different components of FVGen. Our full model presents the most visual similarity between the ground truth.

Assessing student initialization. As it can be seen from the first column of Figure 6 and Table 3, student initialization is a pivotal module of our architecture. Without student initialization with GAN optimization, the student model would fail to generate samples close to the teacher distribution, thus the teacher score function is not reliable for distillation.

Assessing distribution matching. Optimization of DMD continues to align the student model distribution with teacher model. Comparing with full model, although the perceptual metrics are not significantly degraded, qualitative results show that without distribution matching, the generated results appear significant blurriness and artifacts.

Assessing soften reverse KL-divergence. As discussed in Section 4.2, soften reverse KL-divergence enables more stable and robust training. Therefore, the DMD optimization achieves better convergence and generated results are more distilled into teacher distribution.

6. Limitations

We identify several limitations in our current work. First, although FVGen achieves similar visual quality compared to ViewCrafter, it is limited by the drawbacks of ViewCrafter: e.g, degradation in structural integrity and consistency when experiencing extremely sparse inputs. Second, FVGen benefits from the synergy of three video diffusion models; therefore, due to computational constraints, we were only able to train the model with videos up to only 16 frames, which may not provide enough coverage for extremely large scenes. Therefore, our future work would address on the limitations mentioned above.

7. Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0075. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. Technical report, 2021. 2, 6, 7, 8

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[3] Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Nitrofusion: High-fidelity single-step diffusion through dynamic adversarial training. *arXiv preprint arXiv:2412.02030*, 2024. 3

[4] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 3

[5] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 7

[6] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. 2024. 2, 3, 5, 6, 7

[7] 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. 2, 3

[8] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2

[9] Dibya Ghosh. Kl divergence - intuition and examples, n.d. Accessed: 2025-03-07. 5

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[12] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6

[14] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 3

[15] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, pages 428–447. Springer, 2024. 3

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2

[17] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 2, 6, 7, 8

[19] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. *arXiv preprint arXiv:2402.03908*, 2024. 2

[20] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. *arXiv preprint arXiv:2406.17601*, 2024. 2

[21] 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. 2

[22] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 2, 3

[23] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025. 2, 4

[24] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 2, 6

[25] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 2, 3

[26] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Yihong Luo, Xiaolong Chen, Xinghua Qu, Tianyang Hu, and Jing Tang. You only sample once: Taming one-step text-to-image synthesis by self-cooperative diffusion gans. *arXiv preprint arXiv:2403.12931*, 2024. 3

[28] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

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

[31] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[33] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[34] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 2, 3

[35] 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. 2

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[38] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 3

[39] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[40] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3, 6

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

[42] 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. 2, 3

[43] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 2

[44] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2, 3

[45] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 7

[46] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8196–8206, 2024. 2, 3

[47] Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025. 5

[48] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[49] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 2, 3, 4, 5, 8

[50] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. *arXiv preprint arXiv:2412.07772*, 2024. 2, 3, 4, 6, 8

[51] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems*, 37: 47455–47487, 2025. 2, 3, 4, 5, 6, 8

[52] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

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