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.

DOLLAR: Few-Step Video Generation via Distillation and Latent Reward Optimization

Zihan Ding*, Chi Jin
Princeton University
{zihand, chij}@princeton.edu

Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, Yuchen Liu
Adobe Research
{diliu, hazheng, krishsin, qiangz, yankang, zlin, yuliu}@adobe.com



Figure 1. By incorporating variational score distillation, consistency distillation and latent reward fine-tuning, our DOLLAR method with 4-inference steps increases sample diversity and fidelity over DMD [75], generates higher quality videos than LCM [36, 62], and achieves $\times 15.6$ acceleration compared with teacher. You can [*click*] rightmost column samples to play videos (20 extracted frames) in Adobe Acrobat. More in Appendix 11.

Abstract

Diffusion probabilistic models have shown significant progress in video generation; however, their computational efficiency is limited by the large number of sampling steps required. Reducing sampling steps often compromises video quality or generation diversity. In this work, we introduce a distillation method that combines variational score distillation and consistency distillation to achieve few-step video generation, maintaining both high quality and diversity. We also propose a latent reward model fine-tuning approach to further enhance video generation performance according to any specified reward metric. This approach reduces memory usage and does not require the reward to be differentiable. Our method demonstrates state-of-the-art performance in few-step generation for 10-second videos (128 frames at 12 FPS). The distilled student model achieves a score of 82.57 on VBench, surpassing the teacher model as well as baseline models Gen-3 [10], T2V-Turbo [29], and Kling [27]. One-step distillation accelerates the teacher model's diffusion sampling by up to 278.6 times, enabling near real-time generation. Human evaluations further validate the superior performance of our 4-step student models compared to teacher model using 50-step DDIM sampling. Project page: https://quantumiracle.github.io/dollar/

1. Introduction

Diffusion probabilistic models [17, 53, 55, 56] have recently revolutionized generative modeling in continuous domains. With remarkable expressive power and flexibility across diverse data formats and modalities, diffusion models have significant breakthroughs in tasks such as text-to-image and text-to-video (T2V) generation. However, despite substantial improvements

*Work done during internship at Adobe Research.

in generation quality, the efficiency of diffusion models remains a limiting factor in practical applications due to the inherently large number of iterative sampling steps. This efficiency challenge is exacerbated in video generative modeling, where the higher-dimensional space demands larger model sizes, more extensive training data, larger input and output tensors, and more sampling iterations. Furthermore, practical applications often require generation qualities that may differ from the training distribution—such as higher aesthetic standards or diverse stylistic choices—necessitating efficient post-training adjustments or fine-tuning to meet specific requirements while managing the substantial cost of pre-training.

To address the efficiency challenges in diffusion models, model distillation [35, 40, 47] has been widely researched across various models and domains. Score distillation, specifically, aims to improve efficiency in 3D [43, 60, 66] and image synthesis [37, 48, 71, 75] by aligning the distribution between teacher and student diffusion models. However, despite achieving high sample fidelity, it often encounters model collapse issues [34, 75]. Another approach, consistency distillation (CD) [57], seeks to ensure consistent sample predictions along the diffusion trajectory. While CD promotes greater sample diversity, it has limitations: it tends to lower sample fidelity and can produce overly smooth outputs in large-scale T2V applications.

A further challenge with distillation methods is that the student's performance is typically upper-bounded by the teacher model. Previous efforts to address this limitation have involved integrating variational score distillation (VSD) [66, 74] or consistency distillation [23] with GAN [13] loss, which modestly enhances sample fidelity within the training distribution, limited by the sparsity of discriminative signals of adversarial training. Consequently, the generated samples may still fall short in capturing nuanced visual quality details and text-to-image alignment, both of which require denser feedback signals. Recently developed image or video reward models offer promising potential to address this gap, providing richer signals for fine-grained improvements in generation quality.

In this work, we address the limitations of consistency distillation (CD) by incorporating a larger number of teacher denoising steps and combining CD with variational score distillation (VSD) to produce high-quality, diverse samples with a few-step model after distillation. However, this alone does not suffice to outperform the teacher model or reliably meet specific preferences for downstream applications, as generated samples may still face challenges in visual quality and text-to-video alignment. While model post-training with a high-quality dataset is a potential solution, it is often costly to implement. To overcome these limitations, we further introduce an efficient reward model fine-tuning method that enhances the student model beyond the teacher's capabilities and aligns it with any pre-defined requirements through tailored reward metrics. The improved performance is shown in Fig. 1.

We propose learning a dual reward model within the latent space, guided by the pixel-space reward model, and utilize the gradients from this latent reward model (LRM) to fine-tune the diffusion model directly. This approach combines the strengths of reward-gradient methods in pixel space and stochastic policy gradient methods, offering several advantages: (1) it harnesses the rich gradient information from the latent reward model, enabling efficient and effective tuning; (2) it does not require the original reward model to be differentiable, broadening applicability to a variety of reward models; (3) it significantly reduces computational and memory costs during fine-tuning by eliminating the need for backpropagation through large pixel-space reward models and the decoder. The LRM approach is versatile, accommodating various reward types—including image, video, text-image, and text-video rewards—thereby enhancing practical usability.

In summary, our contributions are threefold: (1) We introduce a diffusion model distillation method that combines VSD and CD losses to enable efficient, few-step T2V models; (2) We enhance CD with a generalized approach incorporating multiple teacher denoising steps to improve its effectiveness; (3) We propose to use a compact latent-space reward model for reward-based fine-tuning, which posts no requirement on the differentiability of original reward metrics and is more memory- and computation-efficient. All evaluations are conducted on large-scale T2V settings. Putting together these innovations, we present `DOLLAR` method with **D**istil**l**ation and **La**tent **R**eward **O**ptimization, to significantly advance the quality and efficiency of video generation and pave the way for real-time applications.

2. Related Work

Video Generation. Recent advancements have extended diffusion models from image synthesis to video generation, addressing the complexities of spatiotemporal data. Pioneering works [18] adapted diffusion processes to handle temporal dynamics, enabling high-fidelity video generation. To enhance computational efficiency, latent diffusion models (LDM) [46] perform diffusion modeling in compressed latent spaces, a strategy further refined for video by [4–6, 10, 14, 18, 19, 22, 52, 58, 61, 70, 73], which generate videos conditioned on textual descriptions, as known as the text-to-video (T2V) models. Additional recent methods [2, 21, 63, 77] have further enhanced video quality. The diffusion transformer (DiT) [38] introduced a scalable diffusion approach by integrating the transformer architecture, supporting training across videos with diverse resolutions and durations [42]. Despite these advancements, challenges like computational cost, temporal consistency, and suitable evaluation metrics persist, guiding future research in diffusion model-based video generation.

Efficiency of Diffusion Models. Diffusion models have achieved state-of-the-art performance in generative tasks, though they are computationally demanding, often requiring hundreds of sampling steps [17, 54, 55]. To improve efficiency,

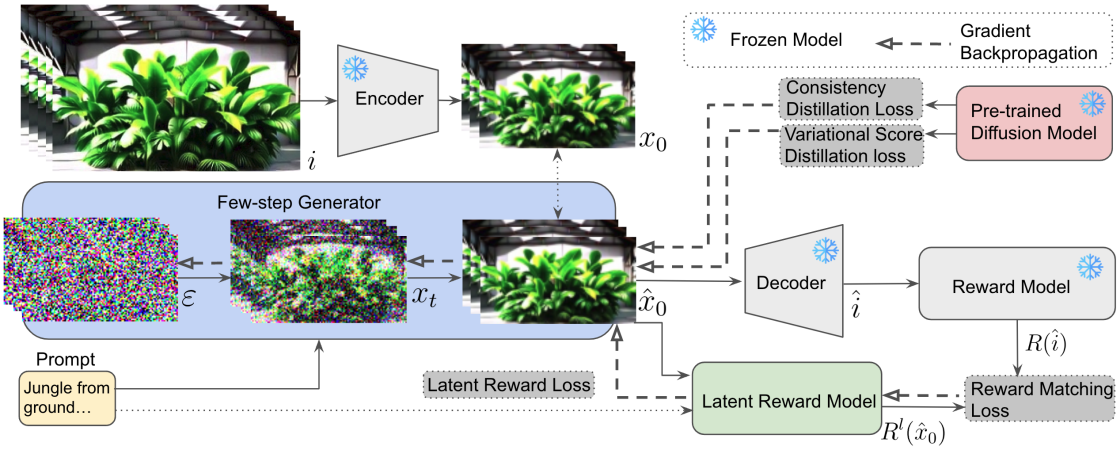


Figure 2. Method Overview: The few-step generator G_θ is trained to generate high-quality samples from random noise in latent space, guided by a combination of variational score distillation (VSD), consistency distillation (CD), and latent reward model (LRM) fine-tuning objectives. VSD loss enhances sample quality, albeit with a risk of mode collapse, while CD loss increases sample diversity without compromising generation quality. The LRM enables reward-based optimization to further improve sample quality, by bypassing the large, pixel-space reward model and the decoder, thereby reducing memory usage and removing the need for differentiable reward models.

distillation methods [35, 47] have been developed to accelerate sampling. With a distribution matching objective between teacher and student models, score distillation was first introduced for 3D generation with diffusion models [43], followed by variational score distillation (VSD) applied to both 3D [66] and image generation [37, 48, 71, 75]. Adversarial training combined with diffusion models has also been proposed to improve few-step image generation [49, 50, 69, 74]. Another line of methods imposes alternative constraints on diffusion trajectories. Consistency models [57] distill diffusion models using consistency distillation (CD) loss, advancing image [36] and video generation like VideoLCM [62], AnimateLCM [59] and FastVideo [76]. Rectified flow methods [31–33] encourage a straight denoising trajectory during training to enable high-quality, few-step sampling. However, existing methods have limitations: VSD-based approaches like distribution matching distillation (DMD) [75] can generate high fidelity samples, but often suffer from model collapse, producing less diverse samples post-distillation. Consistency models tend to yield lower fidelity and samples that are qualitatively poorer than those of the teacher model, though the sample diversity is preserved. Our distillation method, which combines VSD and CD, addresses these issues by generating high-quality and diverse samples.

Reward-based Fine-tuning. To further enhance image and video generation quality in aspects like visual quality and text-image alignment, researchers have introduced various reward-based fine-tuning techniques for diffusion models [3, 7, 8, 8, 9, 28, 29, 44, 72]. Direct reward gradient methods, such as diffusion posterior sampling [7], ReFL [72], and DRaFT-K [8], require a differentiable reward model in pixel space. Approaches like VADER [44] and T2V-Turbo [30] extend these methods to video models.

In contrast, denoising diffusion policy optimization (DDPO) [3] conceptualizes the denoising process as a decision-making problem, employing stochastic policy gradient algorithms such as REINFORCE and PPO for optimization without the need for a differentiable reward function. However, DDPO has been observed to be less sample-efficient compared to reward-gradient methods due to the lack of direct gradient information [8]. To utilize the rich reward gradient information without the computational overhead of backpropagating through large reward models and decoders, we propose using a latent reward model as a surrogate for pixel-space reward models in video synthesis. This approach enables efficient gradient-based fine-tuning of diffusion models within the latent space.

3. Methodology

The overview of our method is shown in Fig. 2.

3.1. Diffusion Model

Suppose the data distribution is $x_0 \sim q(x_0)$, the diffusion model approximates this distribution by gradually denoising along a Markov chain. Forward diffusion follows $x_t := F(x_0,t) = a_t x_0 + b_t \varepsilon, \varepsilon \sim \mathcal{N}(0,\mathbf{I})$. For DDPM,

$$a_t = \sqrt{\bar{\alpha}_t}, b_t = \sqrt{1-\bar{\alpha}_t} \tag{1}$$

with $\bar{\alpha}_t = \Pi_{i=1}^{t} \alpha_i$ following a pre-specified noise schedule $\alpha_t, t \in [T]$. For the general variance-preserving schedule [56], it satisfies $a_t^2 + b_t^2 = 1$, therefore it can be equivalently written as $x_t = \cos(t) x_0 + \sin(t)\varepsilon, t \in [0, \frac{\pi}{2}]$ with a simple mapping of time sequences. Standard diffusion model optimization with velocity prediction v_θ follows the loss:

$$\mathcal{L}_V(\theta) = \mathbb{E}_{x_0 \sim q(x_0), \varepsilon \sim \mathcal{N}(0,\mathbf{I}), t}\left[w_t || v_\theta(x_t,t) - v_t ||_2^2\right] \tag{2}$$

$$v_t = -\sin(t) x_0 + \cos(t)\varepsilon \tag{3}$$

For rectified flow (RF) [32] or flow matching [31], $a_t = 1 - t, b_t = t, t \in [0, 1]$, with a constant velocity target $v_t = \varepsilon - x_0, \forall t \in [0,1]$.

After training, the reverse diffusion process follows $x_{t-1} \coloneqq \text{Denoise}(x_t, t, \theta)$ iterative denoising. The training and inference details of our models refer to Appendix 7.2.

3.2. Consistency Distillation

Consistency model [57] enforces the consistency loss as the distillation method from a pre-trained teacher model $v_{\theta'}$, with a discrete sub-sampled time schedule $t_1 = \epsilon < t_2 < \cdots < t_N = T$:

$$\mathcal{L}_{\text{CD}}(\theta) = \mathbb{E}_{x_0 \sim q(x_0), t_n}[\lambda(t_n) d(f_\theta(x_{t_{n+m}}, t_{n+m}), f_{\theta^-}(\hat{x}_{t_n}, t_n)] \tag{4}$$

$$\hat{x}_{t_n} = \text{Denoise}^m(x_{t_{n+m}}, t_{n+m}, t_n, \theta') \tag{5}$$

where $\lambda(t_n)$ is a time dependent coefficient usually set as constant in practice, $d(\cdot, \cdot)$ is a distance metric like MSE or Huber loss, and θ^- is exponential moving average of θ. Instead of traditional one-step denoising with the teacher model, we apply a generalized CD with $\text{Denoise}^m(\cdot)$ indicating the m-step denoising function as defined by Eq. (14), which iteratively predicts the sequence $(\hat{x}_{t_{n+m-1}}, \ldots, \hat{x}_{t_n} | x_{t_{n+m}})$. This is practically found to improve generation quality. The student consistency function f_θ can be reparameterized from the neural network prediction, similar as in LCM [36]:

$$f_\theta(x_{t_n}, t_n) = c_{\text{skip}} x_{t_n} + c_{\text{out}} x_\theta(x_{t_n}, t_n)$$

To enhance the distillation for conditional generation with conditional variable $c \in \mathcal{C}$ (*e.g.*, text prompts), we applied the classifier-free guidance (CFG) [16] augmentation for the teacher denoising function, similar as VideoLCM [62], but for v_θ-prediction in our case:

$$v_\theta^w(x_{t_n}, t_n, c) = v_\theta(x_{t_n}, t_n, c) + w\big(v_\theta(x_{t_n}, t_n, c) - v_\theta(x_{t_n}, t_n, \varnothing)\big) \tag{6}$$

This is applied in replacement of v_θ in Eq. (15) for conditional generation.

3.3. Variational Score Distillation

Variational score distillation (VSD) [66] is proposed with the objective of distribution matching between the teacher and student models, by approximating the scores with properly trained diffusion models. Specifically, the loss of minimizing the Kullback-Leibler (KL) divergence between real (teacher) sample distribution p_{real} and fake (student) sample distribution p_{fake} has the form:

$$\mathcal{L}_{\text{VSD}} \coloneqq D_{\text{KL}}(p_{\text{fake}} || p_{\text{real}}) = \mathbb{E}_{x \sim p_{\text{fake}}}[\log \frac{p_{\text{fake}}(x)}{p_{\text{real}}(x)}] \tag{7}$$

$$= \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \mathbf{I}), x = G_\theta(\varepsilon)}[\log \frac{p_{\text{fake}}(x)}{p_{\text{real}}(x)}] \tag{8}$$

and the derivative is,

$$\nabla_\theta D_{\text{KL}} = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \mathbf{I}), x = G_\theta(\varepsilon)}[-(s_{\text{real}}(x) - s_{\text{fake}}(x)) \nabla_\theta G_\theta(\varepsilon)] \tag{9}$$

with score functions $s_{\text{real}}(x) = \nabla_x \log p_{\text{real}}(x)$ and $s_{\text{fake}}(x) = \nabla_x \log p_{\text{fake}}(x)$ for two distributions. $G_\theta(\cdot)$ is the generation process by the student network through iteratively denoising the noisy training samples. The real score s_{real} is estimated with the pretrained teacher model.

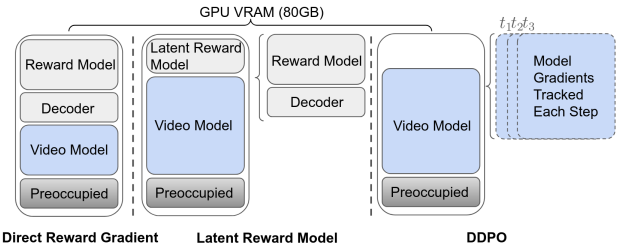


Figure 3. Comparison of different reward fine-tuning methods: (1) Direct reward gradient methods are limited to small reward or video models or short input sequences, and they also require a differentiable reward model; (2) The latent reward model is compact and bypasses the decoder for gradient-based optimization, making it suitable when large reward models and decoders exceed available VRAM; (3) DDPO is similarly constrained by VRAM limits when handling large video models and tracking log-probabilities of samples over multiple steps.

For accurately estimating the fake score s_{fake}, the fake diffusion model $v_{\theta_{\text{fake}}}$ is initialized from the teacher and dynamically adapts according to the student sample distribution. For the score estimation purpose, the fake score model is updated with the diffusion loss $\mathcal{L}_{\text{CV}}(\theta_{\text{fake}})$ following Eq. (13), on student generated samples.

State-of-the-art distillation method distribution matching distillation (DMD) [74, 75] mainly applies the VSD loss for student distillation. We abandon the adversarial loss in [74] since the GAN training is usually unstable and the improvement can be marginal. We replace the adversarial training with the consistency distillation loss and reward model fine-tuning, which generates richer gradient signals as evidenced in our experiments.

3.4. Latent Reward Fine-tuning

Reward fine-tuning is an effective approach to align the sample distribution with the specified preference metric in the post-training phase, as evidenced Fig. 6.

Previous reward-based optimization methods either (1) requires direct gradients from the pixel-space reward models [8, 72], or (2) relies on the log-probability estimation of the samples for multiple diffusion steps like DDPO [3], as compared in Fig. 3. One major drawback of (1) is that it only works for differentiable reward function, while not feasible for non-differentiable ones like JPEG compressibility [3], etc. Apart from that, the reward models usually work for raw RGB pixel space, which requires the reward gradient to backpropagate through not only the large reward models, but also the decoder, as the practical framework usually follows LDM [46] for latent space modeling. Method (2) is found to be less efficient in reward optimization due to lack of rich reward gradient information [8], and occupies more memory due to gradient

estimation over multiple diffusion steps. A concurrent work [28] explores the reward in latent space, but is limited to image generation. Video models have a pressing need for compact reward models, as a largely unexplored area, where leveraging image-based rewards for video generation remains non-trivial.

We propose to learn a dual latent reward model (LRM) for directly optimizing the video diffusion model in the latent space, which supports any type of reward metrics as detailed in Appendix 8.3. Here we take image rewards as an example. Consider a provided image reward model $\mathcal{R}: \mathcal{I} \to \mathbb{R}$ with RGB image $i \in \mathcal{I}$ as its input, we approximate the LRM $\mathcal{R}^l_\phi: \mathcal{X} \bigcup \mathcal{X}' \to \mathbb{R}$ with loss:

$$\mathcal{L}_{\text{LRM}}(\phi) = \mathbb{E}_{x \in \mathcal{X} \bigcup \mathcal{X}'} \left[\left(\mathcal{R}(\text{Dec}(x)) - \mathcal{R}^l_\phi(x) \right)^2 \right] \tag{10}$$

where $\mathcal{X}' = \{G_\theta(\varepsilon)\}$ is the set of generated samples from the generator, and $\text{Dec}(\cdot)$ is the decoder. We use both real images and generated images to improve the robustness of learned LRM on generated samples. To alleviate the computational burden in training the LRM, we apply Eq. (15) for single-step prediction of generated samples, rather than iteratively denoising along the entire trajectory. This approach significantly reduces memory usage by avoiding gradient backpropagation through the iterative sampling process. Although our distilled student models operate with a maximum of 4 sampling steps, memory usage can still be intensive if samples are generated with a full denoising process. Tab. 1 compares the parameter counts and memory costs for pixel-space reward models and LRMs on video samples with a batch size of 1. HPSv2 and PickScore are two pixel-space reward models used in our experiments (as described in Sec. 5.3).

Table 1. Comparison of parameters and GPU memory (VRAM) costs and for pixel-space HPSv2, PickScore reward models and LRMs.

Model	# Parameters	Forward VRAM	Backward VRAM
Image LRM	189,441	8.998 MB	17.772 MB
Text-image LRM	763,009	15.500 MB	26.277 MB
HPSv2/PickScore	632 million	5.926 GB	>90 GB

With the compact and differentiable LRM on the latent space, we apply direct reward gradient optimization to fine-tune the diffusion model:

$$\mathcal{L}_{\text{FT}}(\theta;\phi) = -\mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\mathbf{I})} [\mathcal{R}^l_\phi(G_\theta(\varepsilon))] \tag{11}$$

In practice, we can either pre-train the LRM first and then fine-tune the diffusion model with a fixed LRM, or train the LRM and fine-tune the diffusion model iteratively. For simplicity, we adopt the second approach. If the original reward model is conditioned on additional text input, $\mathcal{R}(i,c)$, the LRM also operates conditionally as $\mathcal{R}^l(x,c), c \in \mathcal{C}$. *The LRM method can accommodate any type of reward models, including image, video, text-image, and text-video rewards.* For image-only LRMs, we use a convolutional neural network, while for text-image LRMs, we apply a cross-attention module after the convolutional feature extractor to integrate text embeddings with image features. For video-based LRMs, the 2D convolution is replaced with a 3D convolutional neural network. Additional details are provided in Appendix 8.3.

4. Multi-Objective Distillation

Distillation using VSD (as in DMD method) alone can result in severe mode collapse, while CD (as in LCM method) tends to produce lower-quality samples by averaging across sample distributions, with evidence shown in Fig. 1 and Fig. 4. By integrating consistency distillation, variational score distillation, and latent reward fine-tuning objectives, our method enables few-step generation of high-quality, diverse samples after distillation, optimized by the following loss function:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{VSD}}(\theta) + \beta_{\text{CD}} \mathcal{L}_{\text{CD}}(\theta) + \beta_{\text{FT}} \mathcal{L}_{\text{FT}}(\theta;\phi) \tag{12}$$

During distillation, the fake score network is updated with $\mathcal{L}_{\text{CV}}(\theta_{\text{fake}})$, and the LRM $\mathcal{R}^l_\phi$ is updated using $\mathcal{L}_{\text{LRM}}(\phi)$. Pseudo-code of our method see Appendix 7.1.

5. Experiments

5.1. Implementation

Student and teacher models. The video diffusion model in our experiments is based on the diffusion transformer (DiT) architecture [38]. Although our methodology is architecture-agnostic and could be applied to models like U-Net [46], we select the transformer due to its scalability. The teacher and student T2V diffusion models are same as a modified variant of Open-Sora [79] and follow the LDM framework [5, 46], utilizing DiT modeling in the latent space encoded with a pretrained 3D variational autoencoder (VAE) [24, 79]. The 3D VAE encodes and decodes videos chunk-by-chunk to alleviate the computational burden, encoding chunks of 16 video frames into 5 latent embeddings. These embeddings are then patchified into sequences as inputs to the DiT. Leveraging the DiT architecture, the model can accommodate arbitrary video durations and resolutions; however, our experiments constrain the video generation setting to 128 frames at a resolution of 192×320, resulting in a patchified sequence of length 9600.

Model training and inference. The teacher model employs DDPM with 1000 sampling steps but uses DDIM for inference with 50 steps, while the student model is distilled to a 4-step sampling similar as previous work [30], with CFG weights 7.5 for CD. The distillation experiment for each student is on 8 A100 GPUs. More implementation details see Appendix 7.3.

Reward Metrics. We utilize Human Preference Score v2 (HPSv2) [67] and PickScore [25] as the text-image reward models. Both are fine-tuned CLIP-type models trained on extensive text-to-image datasets with human preferences.

While our methods are compatible with directly optimizing the model using VBench reward metrics, we intentionally avoid doing so, as VBench scores serve as one of the final evaluation criteria. However, in Appendix 9.3, we present additional experiments where we optimize using VBench metrics, such as dynamics degree, and other image rewards, like JPEG compressibility. These experiments reveal that while reward scores can be significantly improved, it could result in overoptimization for specific metrics, leading to a degradation in overall generation quality. Consequently, we adopt the more general preference-based reward models, HPSv2 and PickScore, by default for reward fine-tuning. Nonetheless, our method remains compatible with other reward models.

Evaluation. To faithfully reflect model performance, we apply both automatic evaluation benchmark VBench [20] and human evaluation for our results. VBench assesses 16 dimensions encompassing both video visual quality and semantic alignment aspects for T2V models, with higher scores indicating better performance in each metric. Following the standard VBench evaluation protocol, we use a set of 946 long prompts, generating five videos per prompt. Final scores for each dimension are averaged across all generated videos for that metric.

Additionally, we examine the impact of prompt length, comparing the performance of long descriptive prompts with short prompts in VBench evaluation (details in Appendix 9.4.1). To further assess text-video alignment capabilities, we sample the distilled student models with various styles and motions, with the results provided in Appendix 11.6.

5.2. Comparison with SOTA Methods and Models

Comparison with SOTA Distillation Methods. In Fig. 4, we compare the samples of our distilled models with SOTA distillation methods including (1). VSD based method for DMD [74, 75]; (2). CD based method typically for LCM [36] and VideoLCM [62], while also applied by T2V-Turbo [30] and FastVideo [76]. These distillation methods apply 4-step inference. Samples by our teacher model with 50 DDIM steps are also displayed. LCM generates blurry samples with lower quality than others, due to the high volatility in training and sensitivity to hyperparameters, although it is possible to achieve better performance with comprehensive hyperparameter search. DMD has lower fidelity and sample diversity compared with ours, with visual evidence in Fig. 27. The LRM on distilled few-step models further enhances quality and alignment, surpassing the teacher model's performance limits. Due to the gradient backpropagation nature of reward gradients, applying them to teacher models with large sampling steps is challenging, whereas they are well-suited for distilled students. More visual results are provided in Appendix 11.

Comparison with SOTA Models on VBench. The VBench evaluation results are summarized in Tab. 2, with the highest

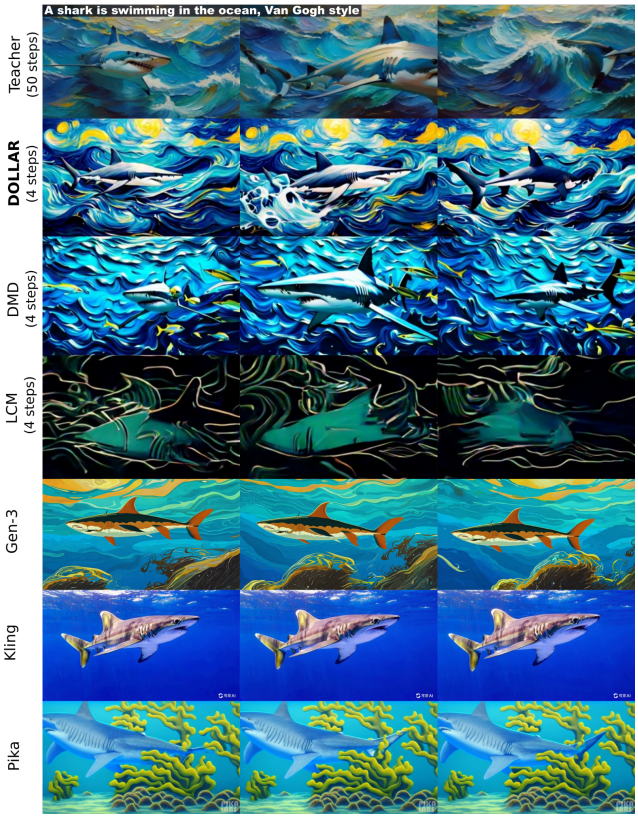


Figure 4. Video samples synthesized with our method and SOTA distillation methods and top video models (no distillation) on VBench.

Table 2. Comparison of VBench scores for different models.

Model	Pika	Gen-2	Gen-3	Kling	T2V-Turbo	Teacher	Our (PickScore)	Our (HPSv2)
Subject C.	96.76	97.61	97.10	**98.33**	96.28	83.99	93.77	92.57
Background C.	**98.95**	97.61	96.62	97.60	97.02	93.78	96.80	96.14
Temporal F.	**99.77**	99.56	98.61	99.30	97.48	96.42	96.30	97.48
Motion S.	99.51	**99.58**	99.23	99.40	97.34	98.09	97.76	98.59
Dynamic D.	37.22	18.89	60.14	61.21	49.17	**99.44**	75.83	81.67
Aesthetic Q.	63.15	**66.96**	63.34	46.94	63.04	61.21	63.80	63.14
Imaging Q.	62.33	67.42	66.82	65.62	**72.49**	63.87	69.40	65.61
Object C.	87.45	90.92	87.81	87.24	**93.96**	85.79	91.63	93.84
Multiple O.	46.69	55.47	53.64	68.05	54.65	52.59	69.71	**72.21**
Human A.	88.00	89.20	96.40	93.40	95.20	**99.60**	99.00	99.00
Color	85.31	89.49	80.90	**89.90**	**89.90**	77.00	77.95	74.78
Spatial R.	65.65	66.91	65.09	**73.03**	38.67	51.40	68.56	68.35
Scene	44.80	48.91	54.57	50.86	**55.58**	49.99	55.06	52.72
T. Style	24.44	24.12	24.71	24.17	25.51	**26.45**	24.64	25.23
A. Style	21.89	24.31	**24.86**	19.62	24.42	24.83	24.45	23.50
Overall C.	25.47	26.17	26.69	26.42	**28.16**	27.89	26.93	26.85
Quality	82.68	82.47	**84.11**	83.39	82.57	81.89	83.49	83.83
Semantic	71.26	73.03	75.17	75.68	72.57	73.71	**77.90**	77.51
Total	80.40	80.58	82.32	81.85	81.01	80.25	82.37	**82.57**

in bold and 2nd and 3rd underlined. Our distilled methods with VSD+CD+LRM achieve superior performance over the baselines including Pika [39], Gen-2 [10], Gen-3 [10], Kling [27] and T2V-Turbo [29] with full inference steps, as well as our teacher model. The visualization is displayed in Fig. 4. The highest semantic scores of our models indicate a significant improvement over baselines for text-video alignment. The quality score, which reflects the visual quality, is heavily affected by the frame consistency metrics like subject consis-

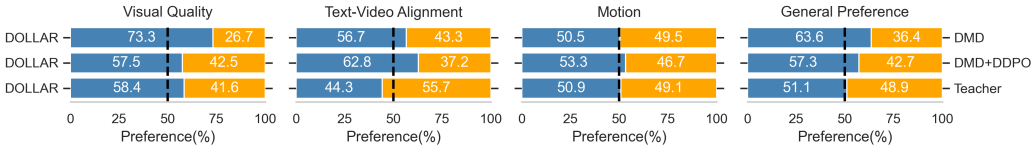


Figure 5. Human preference results for our DOLLAR method (with HPSv2) against baselines. The percentage indicates the preference rate.

tency, background consistency, temporal flickering and motion smoothness, which are usually high if there is a lack of motions in the videos. Our models have significantly higher dynamics degree for motions as shown in the table as well as visualization in Appendix 11. The total score is a weighted sum of all metrics showing the general preference of the videos, and our method achieves 82.37 and 82.57 surpassing all models in the table, as well as outperforming the teacher model. The students achieve higher scores in 9-10 metrics (out of 16) than the teacher. It indicates that the performance of our method is not upper bounded by the teacher model, which is beyond the VSD loss for student and teacher distribution matching. The additional CD loss enforces the self-consistency of model prediction on noisy real images. It provides the source of signals to improve the student model over teacher model on quality and semantic performances, which are further boosted by LRM fine-tuning.

Human Evaluation. We further conduct human evaluation to visually compare the generated videos for different models, over four independent metrics: visual quality, text-video alignment, motion, and general preference. The main evaluation results are shown in Fig. 5, with more details in Appendix 9.2. From the evaluated results, our method with HPSv2 reward is preferred more than the DDPO method (by 57.3%) and teacher model (by 51.1%), and performs similarly with the Gen-3 model (by 45.6%) in terms of general preference. The visual quality of our distilled students is significantly higher than both teacher (by 58.4%) and Gen-3 (by 55.9%). Moreover, we find that, PickScore increases visual quality, but likely leads to worse motion performance. HPSv2 reward tuning not only increases the visual quality, but has better motion and text-video alignment.

5.3. Comparison with Pixel-Space Reward

We compare our LRM with pixel-space reward fine-tuning method DDPO [3], which applies the REINFORCE algorithm to optimize the diffusion model by treating the diffusion process as a MDP. More details refer to Appendix 8.6. Direct reward gradient methods like ReFL [72] and DRaFT [8] exceed single-GPU memory capacity in our case, thus are not included as baselines.

Tab. 5 compares VBench scores and final reward values for LRM and DDPO. The last row "Reward" indicates the corresponding reward value after fine-tuning, for example, PickScore value is reported if the model is fine-tuned with PickScore reward model, and similar for HPSv2. The mean and standard deviation values are reported with 500 videos generated under VBench prompts. As visualized in Fig. 6, the reward fine-tuning

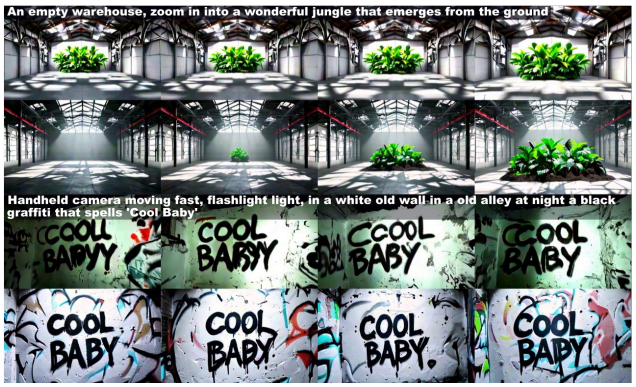


Figure 6. Compare the generated samples with (first line) and without (second line) reward fine-tuning for two samples. First sample: 4 frames are extracted from one sampled video per method along the time sequence. Second sample: one frame is extracted from one video, with 4 videos sampled by the same prompt.

helps to improve the text-image alignment for the first prompt by more explicitly exhibiting the "emerging" effect, and improves the accuracy of text display in frames for the second prompt. The lighting style is also improved through fine-tuning. Full results for VBench scores refer to Appendix 9.5. More details of DDPO and LRM methods including the learning curves for reward optimization are provided in Appendix 8. Human evaluation results of the LRM and DDPO refer to Appendix 9.2.

5.4. Ablation of Distillation Methods

Tab. 3 shows the ablation of our distillation method by comparing it with VSD and VSD+CD. The breakdown results for each VBench metric refer to Appendix 9.5. VSD has comparable performances with teacher, with additional CD loss it increases the sample diversity. Our VSD+CD+LRM method achieves high sample quality and diversity overall. We provide visualization of samples in Appendix 11.

Table 3. Comparison of teacher and students with different distillation methods, with 4-step sampling for student models.

Model	Teacher	Student		
Method	DDIM 50 steps	VSD	VSD+CD	VSD+CD+LRM
Quality Score	81.89	80.95	82.16	**83.83**
Semantic Score	73.71	76.61	74.58	**77.51**
Total Score	80.25	80.08	80.65	**82.57**
Vendi (Pixel)↑	1.46±0.14	1.49±0.14	1.59±0.17	**1.60±0.14**
Vendi (Inception)↑	**2.34±0.16**	1.91±0.14	2.14±0.15	1.98±0.14

Diversity Measure. The diversity of model generation is not captured by the VBench. We conduct both qualitative and

quantitative comparison of generation diversity for different distillation methods. We quantitatively measure the diversity of sampled videos with Vendi score [12], which is based on the similarity matrix for the sample set. The mean and standard deviations across prompted video samples are reported in Tab. 3. Details about the video diversity measure using Vendi score and visual samples are provided in Appendix 9.1. We find that the Inception-based Vendi score aligns better with visual inspection than pixel-based alternatives. While VSD produces high-quality samples, it tends to lead to mode collapse. Our method addresses this diversity limitation by incorporating CD loss, and further enhances generation quality through LRM fine-tuning.

Inference Time. Tab. 4 presents the per-sample inference time consumption for the teacher model using 50-step DDIM inference, and for student models with 1, 2, and 4 inference steps. Here, "diffusion time" refers solely to the diffusion sampling in latent space, while "inference time" encompasses the complete generation process for one video, including text encoding, diffusion sampling, and decoding of latent outputs. The inference experiments are conducted on a single A100 80GB GPU, with mean and standard deviation calculated over 100 samples. With parallel sampling across multiple GPUs, the amortized time per sample can be further minimized. The reported values indicate the percentage of the teacher model's inference time, excluding amortization effects. Distilled student models significantly accelerate diffusion sampling compared to the teacher, achieving speedups from $\times$15.6 (4 steps) to $\times$278.6 (1 step).

Table 4. Time consumption of teacher and distilled student models (as percentage of teacher's total inference time) with different numbers of function evaluations (NFE)

Model	Teacher	Student		
Steps (NFE)	50	4	2	1
Diffusion Time (%)	91.94±0.32	5.88±0.03	2.16±0.01	0.33±0.02
Inference Time (%)	100.00±0.66	13.06±0.17	9.30±0.11	7.45±0.12

Table 5. Comparison of LRM with DDPO using VBench and training reward metrics. DOLLAR is our final method with VSD+CD+LRM.

Reward Model	PickScore			HPSv2		
Method	VSD+DDPO	VSD+LRM	DOLLAR	VSD+DDPO	VSD+LRM	DOLLAR
Quality	82.99	**84.01**	83.49	82.97	83.53	83.83
Semantic	77.26	72.51	**77.90**	74.56	75.67	77.51
Total Score	81.84	81.71	82.37	81.29	81.96	**82.57**
Reward	0.207	0.207	**0.210**	0.271	0.276	**0.277**

5.5. Ablation Studies

Distillation Timesteps. Our proposed method supports an arbitrary subset of timesteps for teacher sampling. By default, we use 4-step sampling for the student model to balance quality and efficiency, as discussed in Sec. 5.1. Here, we investigate the impact of varying the number of sampling steps during distillation, specifically testing 1 step (timestep [999]), 2 steps (timesteps [499, 999]), and 4 steps (timesteps [249, 499, 749, 999]) with equal spacing. While our approach does not require equal spacing, this configuration is used for consistency in this experiment. The evaluated VBench scores are reported in Tab. 6. All three distilled student models with VSD loss demonstrate comparable or even superior performances relative to the teacher model with 50 inference steps. The slight differences can be attributed to checkpoint selection and evaluation variance. From visual inspection and human evaluation, we find that models with more inference steps tend to perform better, which may not be fully captured by the minor differences in VBench scores. Sample visualizations are provided in Appendix 11. The breakdown results for each VBench metric refer to Appendix 9.5.

Model	Teacher	Student (VSD)		
Inference Steps	50	1	2	4
Quality	81.89	81.61	82.71	80.95
Semantic	73.71	76.66	73.86	76.61
Total	80.25	80.62	80.94	80.08

Table 6. Comparison of the number of inference steps for distilled students with VSD using VBench (long prompt).

Model	Teacher	Student (VSD+CD)	
CD w/ Denoise^m	-	m=1	m=5
Quality	81.89	80.75	82.16
Semantic	73.71	71.57	74.58
Total	80.25	78.92	80.65

Table 7. Effect of the number of teacher denoising steps in consistency distillation (CD) on VBench (long prompt).

Consistency Distillation Denoising Steps. In Sec. 3.2, we introduced the consistency distillation method with a multi-step teacher denoising function: Denoise$^m(\cdot)$. We ablate the choice of m in experiments and find that a larger value like $m = 5$ improves distillation performance, as detailed in Tab. 7. The student models follow 4-step schedule, and CD loss is applied on a 50-step DDIM schedule with step size 20 as previously discussed. Full results for VBench scores see Appendix 9.5.

We provide additional ablation studies for the prompt length of VBench evaluation and different student-teacher parameterization in Appendix 9.4.

6. Conclusion and Discussion

We propose the DOLLAR method for diffusion distillation, combining VSD, CD, and LRM objectives to dramatically accelerate teacher inference. This is the first work effectively integrate score matching and self-consistency objectives for large-scale diffusion model distillation, demonstrating its effectiveness in generating 10-second videos. Our experiments validate the feasibility of unifying score matching for fidelity and self-consistency for diversity within a single framework. With this approach, the distilled student models achieve significantly higher VBench scores than the teacher model, enhancing both visual quality and text alignment. Notably, this is accomplished within just 40,000 iterations using 8 GPUs, representing only a small fraction of the training dataset and computational resources required for teacher model pre-training. These results highlight the effectiveness of our method, particularly for reward fine-tuning in latent space. Compared to large pixel-space reward models, LRM is compact and significantly reduces memory costs.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 9

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3, 4, 7, 5, 6, 9

[4] 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. 2

[5] 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. 5

[6] 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. 2

[7] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 3

[8] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 3, 4, 7, 6

[9] Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. *arXiv preprint arXiv:2409.08861*, 2024. 3, 13

[10] Patrick Esser, Robin Rombach, and Björn Ommer. Structure-aware video generation with latent diffusion models. *arXiv preprint arXiv:2303.07332*, 2023. 1, 2, 6

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[12] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 8

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[14] William Harvey, Søren Nørskov, Niklas Kölch, and George Vogiatzis. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022. 2

[15] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 4, 5

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4, 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1, 2

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

[19] Yu Hong, Jing Wei, Xing Liu, Xiaodi Wang, Yutong Bai, Haitao Li, Ming Zhang, and Hao Xu. Cogvideo: Large-scale pretraining for text-to-video generation with transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[20] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6

[21] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 2

[22] Levon Khachatryan, Adrien Davy, Baptiste Emond, and Jun Wang. Text2video-zero: Zero-shot text-to-video generation using pretrained text-to-image diffusion models. *arXiv preprint arXiv:2302.01327*, 2023. 2

[23] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023. 2

[24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[25] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023. 5, 4

[26] Weijie Kong and et al Qi Tian. Hunyuanvideo: A systematic framework for large video generative models, 2024. 2

[27] Kuaishou. Kling. https://kling.kuaishou.com/en, 2024. Accessed: [today's date]. 1, 6

[28] Jiachen Li, Weixi Feng, Wenhu Chen, and William Yang Wang. Reward guided latent consistency distillation. *arXiv preprint arXiv:2403.11027*, 2024. 3, 5

[29] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024. 1, 3, 6

[30] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*, 2024. 3, 5, 6

[31] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3, 4

[32] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4

[33] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 3, 1

[34] Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024. 2

[35] Eric Luhman and Tobias Luhman. Knowledge distillation for generative models. *arXiv preprint arXiv:2106.05237*, 2021. 2, 3

[36] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1, 3, 4, 6, 2

[37] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 5

[39] Pika Labs. Pika Labs. https://www.pika.art/. Accessed: September 25, 2023. 6

[40] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018. 2

[41] Adam Polyak and et al Amit Zohar. Movie gen: A cast of media foundation models, 2025. 2

[42] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

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

[44] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024. 3

[45] Allen Z Ren, Justin Lidard, Lars L Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. *arXiv preprint arXiv:2409.00588*, 2024. 6

[46] 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, 4, 5, 1

[47] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2, 3

[48] Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. *arXiv preprint arXiv:2406.04103*, 2024. 2, 3

[49] 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

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

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4, 5

[52] Uriel Singer, Adam Polyak, Eliya Nachmani, Guy Dahan, Eli Shechtman, and Haggai Hacohen. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[53] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015. 1

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1, 2

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 3

[57] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 2, 3, 4

[58] Ruben Villegas, Jiahui Yang, Sergey Tulyakov, Jan Kautz, and Seungjun Hong. Phenaki: Variable length video generation from open domain textual descriptions. *arXiv preprint arXiv:2210.02399*, 2022. 2

[59] Fu-Yun Wang, Zhaoyang Huang, Xiaoyu Shi, Weikang Bian, Guanglu Song, Yu Liu, and Hongsheng Li. Animatelcm: Accelerating the animation of personalized diffusion models and adapters with decoupled consistency learning. *arXiv preprint arXiv:2402.00769*, 2024. 3

[60] 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

[61] 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. 2

[62] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023. 1, 3, 4, 6

[63] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2

[64] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 4, 5

[65] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 4, 5

[66] 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, 4

[67] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 5, 3, 4

[68] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 4, 5

[69] Tianyu Xiao, Dara Bahri, Pawel Lucjan Stanczuk, Duygu Ceylan, Julian McAuley, Arash Vahdat, and Jan Kautz. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021. 3

[70] Tong Xiao, Peng Liu, and Yi Yang. Dual diffusion models for high-fidelity video generation. *arXiv preprint arXiv:2301.06513*, 2023. 2

[71] Sirui Xie, Zhisheng Xiao, Diederik P Kingma, Tingbo Hou, Ying Nian Wu, Kevin Patrick Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. *arXiv preprint arXiv:2405.16852*, 2024. 2, 3

[72] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4, 7, 5

[73] 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. 2, 9

[74] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. 2, 3, 4, 6

[75] 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. 1, 2, 3, 4, 6, 24

[76] Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhenghong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025. 3, 6

[77] Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, et al. Sf-v: Single forward video generation model. *arXiv preprint arXiv:2406.04324*, 2024. 2

[78] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 2

[79] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 5