Data-Forcing Distillation: Restoring Diversity and Fidelity in Few-Step Video Generation

Siyi Chen[1] **Shaowei Liu**[2,3] **Yixuan Jia**[1] **Zian Wang**[2] **Huan Ling**[2] **Qing Qu**[1] **Jun Gao**[1,2]

[1]University of Michigan, Ann Arbor, MI, USA [2]NVIDIA [3]University of Illinois Urbana-Champaign, IL, USA

Abstract

Recent progress has shown promise in distilling multi-step video diffusion models into efficient few-step students. Among them, Distribution Matching Distillation (DMD) and its successor DMD2 achieved strong generation quality and fast convergence. However, due to the nature of the reverse Kullback-Leibler (KL) objective, these methods exhibit two persistent failure modes: a substantial drop in sample diversity, and visibly over-saturated outputs that deviate from real-video appearance. In this work, we propose *Data-Forcing Distillation* (DFD), a simple post-training framework that restores diversity and fidelity in DMD *with only a single-line of code change*. At its core is the *teacher score discrepancy* to guide the student toward the real-data distribution, pulling it to missing modes (mitigating mode collapse) and away from problematic modes absent in real data (avoiding over-saturation). We provide an in-depth theoretical analysis of our framework and validate our approach on both text-to-video and image-to-video generation. With *only 100–300 steps* of finetuning, DFD effectively restores diversity and fidelity on both Wan2.1-1.3B and Cosmos-Predict2.5-2B model, resolving the over-saturation artifacts with significantly better video dynamics and appearance, and even outperforms the teacher model. We refer the readers to the supplementary material for full video comparisons. Project page: https://csy2077.github.io/dfd-project-page/

1. Introduction

Recent progress in large-scale diffusion and flow-based models has substantially advanced video generation (Song et al., 2020; Lipman et al., 2022; Rombach et al., 2022; Wan et al., 2025; Ali et al., 2025; Kong et al., 2024; Yang et al., 2024; Brooks et al., 2024). Modern video generators can synthesize photorealistic, high-fidelity, and diverse videos from a text prompt or a single image. Yet, this capability comes at a heavy computational cost: high-quality generation typically requires multiple denoising steps, resulting in high inference latency and limited deployment in interactive, real-time, or large-scale production settings. Accelerating diffusion sampling while preserving the quality and diversity of the original multi-step model has thus become a central research challenge.

Two families of distillation methods have emerged to address this challenge. *Trajectory-based* methods train the few-step student to regress the sampling trajectory of a multi-step teacher (Salimans & Ho, 2022; Meng et al., 2023; Song et al., 2023; Song & Dhariwal, 2023; Lu & Song, 2024; Geng et al., 2025; Sabour et al., 2025). While effective at reducing sampling steps, these methods require the student to approximate a complex multi-step trajectory with only a few function evaluations, which becomes challenging for high-dimensional video generation and can result in noticeable quality degradation. *Distribution-based* methods (Yin et al., 2024b;a; Luo et al., 2023b; Wang et al., 2025) relax this trajectory constraint and train the student so that its output distribution matches the teacher's distribution. DMD (Yin et al., 2024b) and its successor DMD2 (Yin et al., 2024a) are representative examples, achieving strong few-step quality and fast convergence; we therefore build on this family in our work. However, when applied to large-scale video generation, distribution-based methods still exhibit two critical failure modes: a substantial drop in sample diversity compared to the teacher model, and visibly over-saturated outputs that deviate from real-video appearance.

The root cause of both failure modes is the reverse Kullback-Leibler (KL) objective used in DMD, which is well known to be *mode-seeking* (Poole et al., 2022; Wang et al., 2023): the student seeks the highest-density modes of p_{real}, as

Correspondence to: Siyi Chen <siyiche@umich.edu>.

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).



Figure 1. **Qualitative results on text-to-video and image-to-video generation**. The colored outline indicates the input image for the image-to-video tasks. Compared to DMD, our method significantly improves visual quality, resolves the over-saturation artifacts with better video dynamics and appearance, and even outperforms the teacher model. The time is measured on a single NVIDIA RTX PRO 6000 GPU.

the reverse KL takes its expectation over student-generated samples (Sec. 3) and primarily optimizes only the modes that the student covers and cannot penalize missing modes in the real distribution. To fix the problem, our intuition is simple: directly incorporating real video samples into the training objective during distillation, so that the student can be explicitly pulled toward the diverse, high-quality real data distribution rather than supervising itself using its own generated data.

Building on this motivation, we propose *Data-Forcing Distillation* (DFD), a video distillation framework that restores both diversity and fidelity with a single line of code change on top of DMD. The core of DFD is the *teacher score discrepancy* that computes the score difference between a real video and the student's generated video, and we directly incorporate it in the distillation objective. Intuitively, this score discrepancy points the student to the real-data distribution, pulling it toward missing modes (mitigating mode collapse) and away from problematic modes absent in real data (avoiding over-saturation). In practice, DFD is implemented with a single-line code change: feeding a real video sample to the teacher in place of the student's own generation when computing the distribution matching objective. While the idea of leveraging real data was already explored in DMD2 (Yin et al., 2024a), it was implemented through an auxiliary GAN (Goodfellow et al., 2014), which destabilizes training and only *implicitly* transfers real-data information to the generator through a discriminator. In contrast, DFD *explicitly* provides real-data guidance to the distribution matching objective, preserving the efficiency and simplicity of DMD while mitigating both failure modes.

We conduct extensive experiments on two different video generation tasks with different pretrained video models to demonstrate the effectiveness of DFD. On text-to-video generation, we distill the pretrained Wan2.1-1.3B (Wan et al., 2025) model into a 4-step student model. On image-to-video generation, we apply the same recipe and distill the pretrained Cosmos-Predict2.5-2B (Ali et al., 2025) model. With only 50-100 steps of post-training on top of DMD2, our DFD consistently and significantly improves visual quality, restores diversity across scene composition and camera motion quality of DMD2, and substantially alleviates over-saturation artifacts. We will release our code and model checkpoints to support future research on efficient, high-fidelity, and diverse video generation.

2. Related Work

2.1. Trajectory-Based Methods

Trajectory-based distillation methods aim to accelerate the sampling process by training a few-step student model to match the denoising trajectory of a multi-step teacher model (Song et al., 2023; Luo et al., 2023a; Meng et al., 2023; Li et al., 2024). Among these, Consistency Models (CMs) (Song et al., 2023; Song & Dhariwal, 2023; Geng et al., 2024; 2025; Sabour et al., 2025; Kim et al., 2023) learns a consistency function $f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \rightarrow \boldsymbol{x}_0$ that directly maps any point $\boldsymbol{x}_t$ on the teacher's probability flow ODE trajectory to the initial data point $\boldsymbol{x}_0$. Initially, discrete-time CMs were trained by minimizing the discrepancy between outputs at adjacent anchor points t and $t - \Delta t$ along the teacher trajectory. Recently, continuous-time CMs (Lu & Song, 2024; Kim et al., 2023) offer a clean upgrade by taking the limit as

$\Delta t \to 0$, simplifying the objective to enforce instantaneous self-consistency. Building upon these foundations, a notable work is the rCM (Zheng et al., 2025), which is the first work to scale up continuous-time consistency models to large image and video diffusion models. Despite their significant empirical success and clean mathematical formulations, trajectory-based distillation methods still face open challenges, often exhibiting noticeable performance degradation when scaled up to large pretrained image synthesis or video generation models (Wang et al., 2024).

2.2. Distribution-Matching Methods

Distribution-matching distillation compresses a multi-step teacher into a few-step student by aligning the student's output distribution with the teacher's. DMD (Yin et al., 2024b) minimizes a reverse KL between the two and achieves stable training and strong sample quality; DMD2 (Yin et al., 2024a) augments it with an auxiliary GAN discriminator to mitigate the mode-seeking behavior of reverse KL, but inherits the well-known instability of adversarial training and only injects the real-data signal *implicitly* through the discriminator. These issues are amplified for video diffusion, where the high spatiotemporal dimensionality makes mode collapse and over-saturation considerably more severe.

A parallel line of work generalizes the divergence itself. f-distill (Xu et al., 2025) casts DMD as a special case of integral f-divergence and reports gains from alternatives such as the Jeffreys divergence, while Uni-Instruct (Wang et al., 2025) unifies Score Implicit Matching (Luo et al., 2024) and Score Identity Distillation (Zhou et al., 2024) under a single objective. Diversity-Preserving DMD (Wu et al., 2026) instead couples DMD with trajectory-based distillation to recover modal diversity, though scaling such hybrids to video remains open. Most recently, Transition Matching Distillation (TMD) (Nie et al., 2026) extends the paradigm temporally by matching conditional transition probabilities across sub-intervals of the denoising trajectory, enabling distillation of large-scale video models – yet it still rests on the same reverse-KL backbone and inherits its diversity limitations.

3. Background: Distribution Matching Distillation

We first review Distribution Matching Distillation (DMD), the reverse-KL objective it optimizes, and the mode-collapse issue that motivates our approach.

Distribution Matching Distillation (DMD) (Yin et al., 2024b;a). To distill a pretrained multi-step teacher into a few-step student G_θ, DMD matches the student's output distribution p_{fake} to the teacher's data distribution p_{real} via a KL objective. Since neither density is tractable, DMD reformulates the gradient in terms of *score functions*, which can be approximated by diffusion denoisers. Concretely, for a noise-perturbed sample $\boldsymbol{x}_t = \alpha_t\, G_\theta(\boldsymbol{z}, c) + \sigma_t\, \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the student gradient takes the form

$$\nabla_\theta \mathcal{L}_{\text{DMD}} \;=\; \mathbb{E}_{t,z,\epsilon}\Big[\, w(t)\, \big(\nabla_x \log p_{\text{fake}}(\boldsymbol{x}_t) - \nabla_x \log p_{\text{real}}(\boldsymbol{x}_t)\big)\, \nabla_\theta G_\theta(\boldsymbol{z}, c)\Big]\,, \tag{1}$$

where $\nabla_x \log p_{\text{real}}$ is learned by the frozen teacher denoiser, $\nabla_x \log p_{\text{fake}}$ is learned by an auxiliary denoiser trained *online* on student samples via denoising score matching, and $w(t)$ is a noise-level weighting. Training alternates between updating θ with Eq. (1) and updating the auxiliary denoiser to track the moving p_{fake}.

Reverse Kullback-Leibler (KL) Loss. DMD instantiates Eq. (1) as the gradient of the *reverse* KL,

$$D_{\text{KL}}\big(p_{\text{fake}} \,\|\, p_{\text{real}}\big) \;=\; \mathbb{E}_{\substack{\boldsymbol{x}=G_\theta(\boldsymbol{z},c)\\ \boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}}\left[\log \frac{p_{\text{fake}}(\boldsymbol{x})}{p_{\text{real}}(\boldsymbol{x})}\right], \tag{2}$$

along the diffusion path whose differentiation yields

$$\nabla_\theta D_{\text{KL}}(p_{\text{fake}} \,\|\, p_{\text{real}}) \;=\; \mathbb{E}_{\boldsymbol{z}}\big[\big(\nabla_x \log p_{\text{fake}}(\boldsymbol{x}) - \nabla_x \log p_{\text{real}}(\boldsymbol{x})\big)\, \nabla_\theta G_\theta(\boldsymbol{z}, c)\big]\,, \quad \boldsymbol{x} = G_\theta(\boldsymbol{z}, c). \tag{3}$$

This formulation gives strong few-step quality and fast convergence, but the reverse KL is well known to be *mode-seeking*: the student collapses onto the highest-density modes of p_{real}, producing low-diversity, often over-saturated samples (He et al., 2024; Lu et al., 2025). The core reason for this mode-seeking behavior is that reverse KL divergence is evaluated as an expectation over the student's generated distribution. Consequently, if the student drops a mode, the loss function incurs no penalty in that region, yielding zero gradient signal to recover the missing mode. The core reason for mode-seeking is that reverse KL samples from the data generated by the student, and when there is a missing mode in the student, the KL divergence can not pull it back.

Regularization with Real Data To overcome the stated mode collapse, over-saturated problem introduced by reverse KL loss, DMD2 (Yin et al., 2024a) partially compensates with an auxiliary GAN (Goodfellow et al., 2014) discriminator, but adversarial training is unstable and only *implicitly* injects the real-data signal through the discriminator rather than through the distribution-matching gradient itself.

4. Data-Forcing Distillation

Our core intuition is to incorporate real data directly into the reverse KL gradient through a differentiable regularizer, where the student's distribution can be explicitly pulled toward the diverse, high-quality real data distribution rather than supervising itself in the teacher. In this section, we first formally formulate our proposed data-forcing distillation, then give a practical implementation.

4.1. The Data-Forcing Distillation (DFD) Framework

Our DFD injects a real-data regularization term into the original DMD gradient.

$$g_{\mathrm{DFD}}(\theta) \;=\; \underbrace{\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\Big[\big(\nabla_x \log p_{\mathrm{fake}}(G_\theta(\boldsymbol{z},c)) - \nabla_x \log p_{\mathrm{real}}(G_\theta(\boldsymbol{z},c))\big)\,\nabla_\theta G_\theta(\boldsymbol{z},c)\Big]}_{\text{native DMD gradient } g_{\mathrm{DMD}}(\theta)} \\ -\; \underbrace{\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{real}}(\cdot|c),\,\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\Big[\Delta_{p_{\mathrm{real}},\,\boldsymbol{x},\,G_\theta(\boldsymbol{z})}\,\nabla_\theta G_\theta(\boldsymbol{z},c)\Big]}_{\text{real-data regularizer}}. \tag{4}$$

$$\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})} \;=\; \nabla_x \log p_{\mathrm{real}}(\boldsymbol{x}) - \nabla_x \log p_{\mathrm{real}}(G_\theta(\boldsymbol{z},c)), \quad \boldsymbol{x}\sim p_{\mathrm{real}}(\cdot\mid c). \tag{5}$$

We refer to $\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}$ as the **teacher score discrepancy**: it measures the gap, in the teacher's score field, between a real sample and the student's generation. When the student matches the real distribution ($p_{\mathrm{fake}} = p_{\mathrm{real}}$), this discrepancy is minimized to zero in expectation, $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}] = 0$ (proof in the Appendix A). Empirically, drawing $\boldsymbol{x}$ from diverse, high-quality real-world data and minimizing $\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}$ pulls the student toward modes it has missed in the real data (mitigating mode collapse) and away from problematic modes such as over-saturated outputs that are not in the real data. More importantly, properties of real data, such as temporal coherence, physical plausibility, photorealism, can be explicitly distilled into the student through this term.

Discussion: No Need for a GAN. DMD2 (Yin et al., 2024a) provides *implicit* real-data supervision through an auxiliary GAN loss; in contrast, DFD provides *explicit* real-data supervision through the score-discrepancy term (Eq. 5). As a result, we drop the GAN loss completely from our distillation pipeline and it will not affect our performance. We validate this in the ablation study in Section 5.2.

4.2. The Underlying Assumption of DFD

The teacher score discrepancy $\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}$ is well-behaved *in expectation*. However, what governs whether this regularizer helps or hurts training is not its expectation but its *variance*—a zero-mean signal can still produce a high-noisy gradient that destabilizes optimization. We therefore need $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[||\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}||^2\right]$ to stay small. To see what controls this variance, a first-order expansion of the teacher's score around $G_\theta(z,c)$ shows that, to leading order, the per-sample discrepancy is approximately proportional to the gap between paired real and generated samples,

$$\left\|\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}\right\| = \left\|\nabla_x \log p_{\mathrm{real}}(\boldsymbol{x}) - \nabla_x \log p_{\mathrm{real}}(G_\theta(\boldsymbol{z},c))\right\| \;\approx\; \alpha_t\, L(c,t)\, \|\boldsymbol{x} - G_\theta(\boldsymbol{z},c)\| \tag{6}$$

where $\boldsymbol{x}\sim p_{\mathrm{real}}(\cdot\mid c)$, α_t is the diffusion schedule coefficient, and $L(c,t)$ is the Lipschitz constant of the teacher's score network in practice. Consequently the variance of the regularizer is approximately proportional to the squared gap $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\|\boldsymbol{x} - G_\theta(\boldsymbol{z},c)\|^2]$, and controlling $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[||\Delta_{p_{real},\boldsymbol{x},G_\theta(\boldsymbol{z})}||^2\right]$ reduces to controlling this gap. We formalize this requirement as the *validity condition*:

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\Big[\|\boldsymbol{x} - G_\theta(\boldsymbol{z},c)\|^2 \;\Big|\; c\Big] \;\le\; \delta(c)^2, \tag{7}$$

where $\delta(c)$ is a conditioning-dependent bound. Combining the approximation above with Eq. 6 yields the corresponding (approximate) control on the regularizer's variance: $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\big[\big\|\Delta_{p_{\text{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}\big\|_2^2 \,\big|\, c\big] \lesssim \alpha_t^2\, L(c,t)^2\, \delta(c)^2$. The complete derivation is in Appendix. A. Since we evaluate the score along the full diffusion path rather than only at its endpoint with pure noise, neither α_t nor $L(c,t)$ will explode.

4.3. Practical Implementation

Eq. (4) admits a clean simplification. Because the regularizer $\Delta_{p_{\text{real}},\,\boldsymbol{x},\,G_\theta(\boldsymbol{z})}$ is evaluated at the same noise level and condition c as the DMD gradient, it exactly cancels the score $\nabla_x \log p_{\text{real}}(G_\theta(\boldsymbol{z}, c))$. The DFD gradient therefore reduces to

$$g_{\text{DFD}}(\theta) \;=\; \mathbb{E}_{\substack{\boldsymbol{x}\sim p_{\text{real}}(\cdot|c)\\ \boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}}\Big[\big(\nabla_x \log p_{\text{fake}}(G_\theta(\boldsymbol{z},c)) - \nabla_x \log p_{\text{real}}(\boldsymbol{x})\big)\,\nabla_\theta G_\theta(\boldsymbol{z},c)\Big]\,. \tag{8}$$

The only difference compared to the DMD gradient (Eq. 3) is the second score: instead of evaluating $\nabla_x \log p_{\text{real}}$ at the student's own output $G_\theta(\boldsymbol{z}, c)$, we evaluate it at a real data sample $\boldsymbol{x}$ drawn under the same condition. Fig. 2 illustrates this difference.

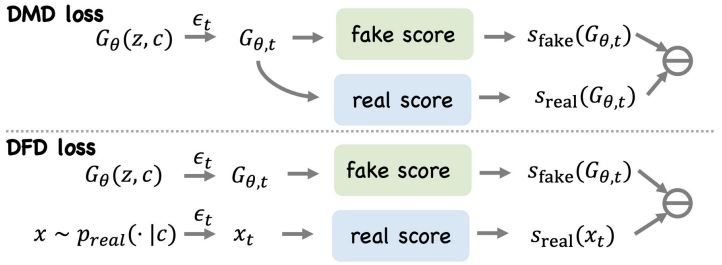


Figure 2. **The comparison between our DFD and the original DMD**. Our DFD computes the real score directly using the videos sampled from the real data distribution, while the original DMD computes the real score using the generated videos from the student.

In practice, we do not enforce Eq. (7) as a hard constraint. Instead, we satisfy it implicitly by applying DFD as a post-training stage on top of a DMD2-pretrained model, whose generations are already close to the real video and naturally keep $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\|\boldsymbol{x} - G_\theta(\boldsymbol{z}, c)\|^2]$ small. The ablation in Section 5.2 confirms the importance of this regime: violating the validity condition—e.g., applying DFD from scratch when the student still produces noisy videos—prevents the model from converging. Concretely, we form the practical update as a combination of the DMD gradient and the DFD gradient,

$$g(\theta) \;=\; (1-w)\, g_{\text{DMD}}(\theta) \;+\; w\, g_{\text{DFD}}(\theta), \tag{9}$$

where $w \in [0, 1]$ controls how much of the DMD signal is retained. We use $w = \frac{1}{2}$ as our default, which preserves the fast convergence of the DMD while leveraging real-data guidance. In practice, we implement Eq. 9 via per-step stochastic sampling, which matches in expectation when $p = w$:

$$\nabla_\theta \mathcal{L} = \begin{cases} g_{\text{DFD}}(\theta) & \text{with probability } p, \\ g_{\text{DMD}}(\theta) & \text{with probability } 1-p. \end{cases} \tag{10}$$

Additionally, following standard practice in the DMD line of work, we replace the exact score in Eq. 9 with the scores estimated by the diffusion models on perturbed samples, and take the expectation over the diffusion timesteps.

Pseudo-Code We provide the pseudo-code of our DFD below, and highlight the simplicity of our method, which only adds one line of code change compared with the original DMD2 method:

4.4. Data Selection for DFD

Once real data enters the distillation gradient, its quality and diversity become the new bottleneck on what the student can learn. We therefore curate the training videos with high quality and diversity. Specifically, we start from the public ViPE-Wild-1M dataset (Huang et al., 2025a), generated by the Wan 2.1 (Wan et al., 2025) 14B model, and cluster its 1M videos into 1,000 clusters. We then manually retain the clusters whose videos exhibit clear semantics, temporal coherence, varied styles, and diverse dynamics. From the retained pool, we construct two training sets: a 20,000-video animation and cartoon set and a complete 30,000-video mixed-style set. We detail the exact selection process in Appendix B

Algorithm 1 Pseudocode of Student Update Step in a PyTorch-like Style.

```
# student_network / teacher_network / fake_score_network: networks
# input_student, t_student: noisy input and timestep fed to student
# t, eps: diffusion timestep and noise for forward diffusion
# data: real data batch; condition: video gen conditioning
# p: probability of using DFD vs DMD
def student_update_step(input_student, t_student, t, eps, data, condition=None):
    # Generate samples using student network.
    gen_data = student_network(input_student, t_student, condition=condition)

    # teacher_data = gen_data.detach() # Original DMD update
    teacher_data = data.detach() if (torch.rand() < p) else gen_data.detach()

    # Inject noise into the data via forward diffusion
    perturbed_data         = forward_diffusion(gen_data, eps, t)
    perturbed_teacher_data = forward_diffusion(teacher_data, eps, t)
    # Estimate scores
    fake_score    = fake_score_network(perturbed_data, t, condition=condition)
    teacher_score = teacher_network(perturbed_teacher_data, t, condition=condition)
    # Compute gradient and student loss
    vsd_grad      = fake_score - teacher_score
    pseudo_target = gen_data - vsd_grad
    gen_loss      = 0.5 * F.mse_loss(gen_data, pseudo_target.detach())
    return gen_loss
```

The violet block is the *only* difference between the DFD and DMD frameworks.

5. Experiments and Results

We evaluate DFD on three representative large-scale video-generation settings: Wan2.1-1.3B for text-to-video generation and Cosmos-Predict2.5-2B for image-to-video generation, We further conduct a comprehensive ablation study to demonstrate the effectiveness of our design choices.

5.1. Main Experiments

5.1.1. Text-to-Video Generation

Experimental Settings. For the text-to-video generation, we distill the Wan2.1-1.3B model on both the animation set and the mix-style set. We compare against two baselines: DMD2 (Yin et al., 2024a) and DP-DMD (Wu et al., 2026). For DMD2, we use the implementation and default configurations from the FastGen codebase. For DP-DMD, we follow the original paper and use a diversity anchor step of $K = 5$ with weight 0.05 (Wu et al., 2026). For all the methods, we distill the pretrained Wan model into a four-step student model. We evaluate results along three aspects. **Video quality**: we adopt the metrics from VBench (Huang et al., 2024). **Video diversity**: we adopt the diversity metric from DP-DMD (Wu et al., 2026). **Camera-pose diversity**: we use the ViPE (Huang et al., 2025a) to estimate the camera poses of each generated video, and compute diversity statistics from the estimated cameras. The evaluation set comprises 70 animation prompts and 70 mix-style prompts, respectively. We generate 8 videos for each prompt with random seeds 1–8, yielding 1120 videos in total for evaluation. Further details are provided in the Appendix B.

Table 1. Quantitative results on the text-to-video experiments using Wan2.1-1.3B. For all metrics, higher is better. The best result among the distilled methods in each column is **bolded**.

	Video Quality						Video Diversity				Camera Pose Diversity	
Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Aesthetic Quality	Average VBench	CLIP (Mean)	CLIP (Per-frame)	DINO (Mean)	DINO (Per-frame)	Endpoint Distance	Trajectory Distance
Teacher	0.956	0.959	0.976	0.985	0.622	0.899	0.178	0.222	0.301	0.350	30.571	26.651
DMD2	0.956	**0.957**	0.973	0.985	0.634	0.901	0.120	0.165	0.190	0.239	9.148	4.284
DP-DMD	**0.960**	**0.957**	**0.977**	0.987	0.633	0.903	0.126	0.165	0.197	0.239	7.208	3.466
Ours	0.956	0.955	0.976	**0.988**	**0.655**	**0.906**	**0.128**	**0.170**	**0.205**	**0.252**	**18.513**	**19.256**



Figure 3. **Qualitative results on text-to-video generation.** The left columns show models distilled on animation set, and the right column on the mix-style set. Our method produces videos that are not over-saturated, and recovers finer details such as the wing texture of the phoenix.

Experimental Results We provide qualitative results in Fig. 3 with quantitative results averaged on the models distilled on the animation test set and the mix-style testset in Table 1. Qualitatively, our method produces higher-quality videos: it mitigates the over-saturation artifacts with significantly better appearance and dynamics, and recovers fine-grained details. Quantitatively, DFD obtains the best overall VBench score among all the distilled models, mainly driven by improved aesthetic quality and motion smoothness. It also improves all four visual-diversity metrics and substantially increases camera-pose diversity compared with DMD2 and DP-DMD. Additional qualitative results are provided in Appendix C.1.

5.1.2. Image-to-Video Generation

Experimental Settings. For image-to-video generation, we distill the Cosmos-Predict2.5-2B model on our curated mix-style dataset. Same as text-to-video generation, all the methods are implemented in the FastGen codebase. For evaluation, we use two image-to-video test sets: one from VBench, containing 348 images, and one curated from the ViPE-Wild-1M dataset, containing 78 images that are held out from training. We adopt the image-to-video metrics from VBench (Huang et al., 2024) for evaluation. Further details are provided in the Appendix B.

Experimental Results. We provide quantitative results in Table 2 and Table 5 (in the appendix), with qualitative results in Fig 4 and Fig 5. Quantitatively, our method outperforms the baselines on the majority of metrics, and is particularly strong at preserving first-frame conditioning and maintaining temporal coherence. Qualitatively, DP-DMD frequently violates the first-frame constraint, occasionally producing abrupt artifacts that disrupt the whole frame DMD2 yields visible quality improvements over DP-DMD but still struggles to maintain temporal coherence in later frames and often hallucinates structural anomalies such as distorted human figures appearing outside vehicles. In contrast, our model adheres closely to the input frame and remains stable and consistent across all generated frames. Our model also achieves better physical plausibility. In the traffic scene example shown in Fig. 5, DMD2 produces visibly implausible interactions, such as vehicles intersecting. DFD reduces such artifacts in this example and yields more coherent motion. These results support the core intuition of our approach: by incorporating real data into the distillation process, the generator can not only mitigate mode-seeking behavior and over-saturation artifacts, but also capture key properties of real videos, such as first-frame coherence, temporal consistency, and physical realism.

5.1.3. Autoregressive Video Generation

Experimental Settings. For autoregressive video generation, we distill the Wan2.1-1.3B model on our curated mixed-style dataset following the *Self Forcing* framework (Huang et al., 2025b). All methods are implemented within the Self Forcing codebase. Additional implementation details are provided in Appendix B.



Figure 4. **I2V results on the ViPE test set.** The colored outline indicates the input image. Our method produces videos that closely follow the first frame and remain coherent across the full sequence, whereas DMD2 and DP-DMD introduce structural anomalies such as an extra cowboy appearing from nowhere (highlighted by white circles).

Table 2. Quantitative comparison on image-to-video generation, we evaluate on the VBench test image suite using the metrics from VBench. For all metrics, higher is better. The best result in each column is **bolded**. Our method consistently outperforms the baselines on all the metrics, and is even better than the teacher model.

Method	Subject Consistency	Background Consistency	Aesthetic Quality	Temporal Flickering	Motion Smoothness	Image-to-Video Subject	Image-to-Video Background	Average
Teacher	0.9616	0.9648	0.6320	0.9717	0.9905	0.9867	0.9925	0.9285
DMD2	0.9421	0.9621	0.6301	0.9734	0.9886	0.9850	0.9900	0.9245
DP-DMD	0.8457	0.9068	0.5850	0.9603	0.9832	0.9570	0.9520	0.8843
Ours	**0.9613**	**0.9692**	**0.6371**	**0.9759**	**0.9900**	**0.9859**	**0.9930**	**0.9303**

Experimental Results. Qualitative results are presented in Fig. 6. Similar to our image-to-video experiments, Self Forcing with DMD2 struggles to preserve temporal coherence over long horizons and frequently introduces structural artifacts and hallucinations. For example, crab-like figures gradually emerge from the stone scene. In contrast, our method maintains stable content and coherent object structures throughout the generated sequence. Our method also produces videos with stronger physical plausibility. In the train example shown in Fig. 6, DMD2 generates physically inconsistent interactions including disconnected train carriages. DFD substantially reduces such artifacts and yields a more coherent and realistic scene. These observations echo our I2V findings and further support the core intuition behind our approach: by incorporating real data into the distillation process, the generator not only alleviates mode-seeking behavior and over-saturation artifacts, but also encourages the model to capture key characteristics of real videos, such as inter-frame coherence, temporal consistency, and physical realism.

5.2. Ablation Study

Table 3. Comparison between models distilled with batch size 16 and 128, evaluated on the VBench test set.

Batch size	Subject Consistency	Background Consistency	Aesthetic Quality	Temporal Flickering	Motion Smoothness	Image-to-Video Subject	Image-to-Video Background	Average
16	0.9613	**0.9692**	0.6371	**0.9759**	0.9900	0.9859	**0.9930**	0.9303
128	**0.9638**	0.9685	**0.6383**	0.9783	**0.9914**	**0.9878**	0.9929	**0.9316**

Effects of the GAN Loss. We ablate the requirement of GAN loss, and compare our DFD pipeline with and without an additional GAN loss. We experiment on the text-to-video generation on animation set. The VBench quality scores are



Figure 5. **I2V results on the VBench test set.** The colored outline indicates the input image. Our method shows much better visual quality with smooth and physically plausible dynamics under complex scenarios, whereas DP-DMD and DMD2 fail to produce valid videos.

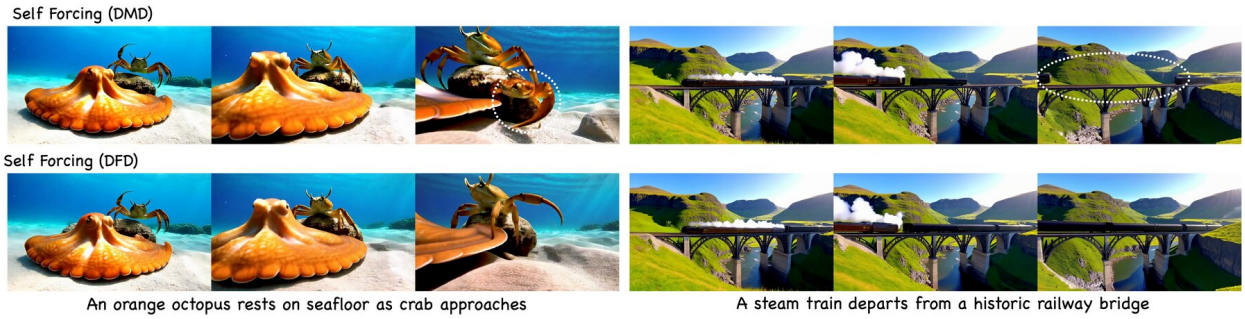


Figure 6. **Autoregressive video generation results.** Results from the generator trained with the Self Forcing pipeline using DFD loss outperform those trained with DMD loss, especially in frame consistency and physical plausibility.



Figure 7. **(a): Ablation on the effect of the GAN loss.** Adding the GAN loss yields no clear quality improvement, and video dynamics even degrade, which is consistent with our quantitative results. **(b): Ablation on the weight in Eq. 9**: $w = 1$ (upper) vs. $w = \frac{1}{2}$ (lower). Each row shows three evenly sampled frames from one generated video. There is no clear visual difference between the two settings. **(c): Qualitative results without DMD2 pretraining.** Initializing purely from the teacher model without DMD2 pretraining fails to produce reasonable results, even after a sufficient number of training steps for both the text-to-video and image-to-video tasks.

reported in Table 6 (in the appendix), and qualitative results in Fig 7 (a). This ablation suggests that, in our post-training setting, adding the GAN loss does not provide a consistent benefit. Removing it simplifies the pipeline and improves dynamic degree and imaging quality, while maintaining comparable scores on the remaining metrics.

Effects on Weights in Eq. 9. We ablate the weights in the Eq. 9. We experiment on the text-to-video generation on animation set. We compare two weighting schemes, $w = \frac{1}{2}$ (our default setting) and $w = 1$ (pure DFD without DMD), on the text-to-video generation. Quantitative results are reported in Table 7 (in the appendix), and qualitative results in Fig 7 (b). The differences between the two settings are small, indicating that our method is stable and insensitive to the choice of w. We adopt $w = \frac{1}{2}$ for theoretical reasons as it better satisfies the condition in Eq. 7.

Effects of DMD2 Pretraining. Our DFD needs to satisfy the *validity condition* in Eq. 7 for stabilized training. To verify this, we ablate by initializing the student with the teacher model (without DMD2 distillation) and comparing the DMD2 initialization. We experiment on both the text-to-video with animation set and image-to-video tasks with mix-style set. Qualitative results are provided in Fig 7 (c). Even after a long training run (e.g., 1400 iterations), the model fails to converge, which underscores the necessity of the validity condition.

5.3. Scaling Up with Large Batch Size

We further scale up the image-to-video generation on the Cosmos-Predict2.5-2B model by increasing the batch size from 16 to 128. As shown quantitatively in Table 3 and visually in Fig 8, the model distilled with a large batch size consistently outperforms the model with a small batch size. The larger batch size yields remarkably better temporal stability during large motions (e.g., the rider's spear) and improved physical plausibility, such as preserving object



Figure 8. **Effect of scaling up the distillation batch size.** The colored outline indicates the input image. Increasing the batch size from 16 to 128 yields videos with superior visual quality and physical consistency. For instance, the spear in the rider's hand exhibits much greater temporal clarity. Additionally, the scaled model preserves exact object counts across frames (the eggs), whereas the smaller batch size struggles with physical plausibility by generating phantom objects out of nowhere (highlighted by the white circle).

permanence (e.g., consistent egg counts).

6. Conclusion

We introduce a data forcing distillation framework that advances the state-of-the-art for few-step video diffusion models, and restores diversity and fidelity. Our DFD requires only a single line of code change to the DMD2 baseline. At its core is the explicit integration of real data into the distribution matching objective, which successfully overcomes the diversity degradation and over-saturation typically induced by reverse-KL formulations. We provide a theoretical analysis to understand our DFD. Through comprehensive, large-scale evaluations on text-to-video and image-to-video generation benchmarks, we demonstrate that our approach successfully resolves persistent failure modes in DMD, yielding substantial improvements in visual fidelity, temporal coherence, and physical plausibility in few-step video generation. We also apply our method to the autoregressive video generation settings under the *Self Forcing* frameworks (Huang et al., 2025b), prelimary results show our method still show the same improvements over the DMD method

7. Limitation

While DFD substantially improves few-step video generation, its performance remains limited under extremely constrained generation budget. As shown in Figure 9, when the generation is restricted to two steps or fewer, the two-step distilled Cosmos-Predict2.5-2B model still struggles to produce high-quality videos: fast-moving objects, such as the man's hands, appear blurry; fine details, such as the woman's facial features, lose fidelity; and the model sometimes collapses to nearly static videos with little motion. These results suggest that, despite the improvements brought by DFD, generating high-quality and temporally dynamic videos remains challenging in the highly aggressive two-step-or-fewer regime.



Figure 9. **Limitation of DFD under a two-step generation budget.** With two-step distillation of the Cosmos-Predict2.5-2B model, DFD still produces blurry results for fast-moving content (e.g., the man's hands), loses fine detail (e.g., the woman's face), or collapses to highly static videos.

References

Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.-W., and other. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Geng, Z., Pokle, A., Luo, W., Lin, J., and Kolter, J. Z. Consistency models made easy. In *The Thirteenth International Conference on Learning Representations*, 2024.

Geng, Z., Deng, M., Bai, X., Kolter, J. Z., and He, K. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

He, J., Chen, W., Zhang, M., Barber, D., and Hernández-Lobato, J. M. Training neural samplers with reverse diffusive kl divergence. *arXiv preprint arXiv:2410.12456*, 2024.

Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.-H., Ren, J., Xie, K., Biswas, J., Leal-Taixe, L., and Fidler, S. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025a.

Huang, X., Li, Z., He, G., Zhou, M., and Shechtman, E. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025b.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Li, J., Feng, W., Fu, T.-J., Wang, X., Basu, S., Chen, W., and Wang, W. Y. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *Advances in neural information processing systems*, 37: 75692–75726, 2024.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Lu, C. and Song, Y. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

Lu, Y., Ren, Y., Xia, X., Lin, S., Wang, X., Xiao, X., Ma, A. J., Xie, X., and Lai, J.-H. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16818–16829, 2025.

Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023a.

Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., and Zhang, Z. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023b.

Luo, W., Huang, Z., Geng, Z., Kolter, J. Z., and Qi, G.-j. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2024.

Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., and Salimans, T. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14297–14306, 2023.

Nie, W., Berner, J., Ma, N., Liu, C., Xie, S., and Vahdat, A. Transition matching distillation for fast video generation. *ArXiv*, abs/2601.09881, 2026. URL https://api.semanticscholar.org/CorpusID:284738120.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sabour, A., Fidler, S., and Kreis, K. Align your flow: Scaling continuous-time flow map distillation. *arXiv preprint arXiv:2506.14603*, 2025.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Song, Y. and Dhariwal, P. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. 2023.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, F.-Y., Huang, Z., Bergman, A. W., Shen, D., Gao, P., Lingelbach, M., Sun, K., Bian, W., Song, G., Liu, Y., et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Wang, Y., Bai, W., Zhang, C., Zhang, D., Luo, W., and Sun, H. Uni-instruct: One-step diffusion model through unified diffusion divergence instruction. *arXiv preprint arXiv:2505.20755*, 2025.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

Wu, T., Li, R., Zhang, L., and Ma, K. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

Xu, Y., Nie, W., and Vahdat, A. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., and Freeman, W. T. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Zheng, K., Wang, Y., Ma, Q., Chen, H., Zhang, J., Balaji, Y., Chen, J., Liu, M.-Y., Zhu, J., and Zhang, Q. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025.

Zhou, M., Zheng, H., Wang, Z., Yin, M., and Huang, H. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024.

A. Additional Theory

A.1. teacher score discrepancy

A good regularizer must not bias the solution: if the student is already perfect, the regularizer should contribute zero gradient. We show this holds for $\Delta_{p_{\text{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}$ in expectation.

Proposition. If the student matches the data distribution exactly ($p_{\text{fake}} = p_{\text{real}}$), then $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\Delta_{p_{\text{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}] = 0$.

Proof. When the student is perfect, $\boldsymbol{x} = G_\theta(\boldsymbol{z})$ is a sample from p_{real}, and $\hat{\boldsymbol{x}}$ is also a sample from p_{real}. They are i.i.d. and therefore *exchangeable*. With shared ϵ,

$$\boldsymbol{x}_t^{\text{fake}} = \alpha_t \boldsymbol{x} + \sigma_t \epsilon \quad \overset{d}{=} \quad \alpha_t \hat{\boldsymbol{x}} + \sigma_t \epsilon = \boldsymbol{x}_t^{\text{real}}. \tag{11}$$

Both have the same distribution $p_{\text{real},t}$, so

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\Delta_{p_{\text{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\nabla_x \log p_{\text{real}}(\boldsymbol{x}_t^{\text{real}})] - \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\nabla_x \log p_{\text{real}}(\boldsymbol{x}_t^{\text{fake}}))] = 0. \tag{12}$$

$\Delta_{p_{\text{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}$ is a zero-mean stochastic perturbation at the optimum. DFD does not bias convergence: it adds a regularizer that vanishes in expectation precisely when the student is correct. □

A.2. Bounded MSE between $g_{\text{DFD}}(\theta)$ and $\nabla_\theta D_{\text{KL}}(p_{\text{fake}} \| p_{\text{real}})$

SETUP

In score-distillation methods, the teacher score is never evaluated on clean inputs. Following standard practice, samples are perturbed by the diffusion forward process to a noise level $t \in (0, T]$ before being passed to the teacher. Let

$$\boldsymbol{x}_t := \alpha_t \boldsymbol{x} + \sigma_t \epsilon, \boldsymbol{x} \sim p_{\text{real}}, \qquad \tilde{\boldsymbol{x}}_t := \alpha_t G_\theta(\boldsymbol{z}, c) + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{13}$$

where (α_t, σ_t) are the diffusion schedule coefficients. The relevant scores are those of the noised marginals,

$$s_{\text{real}}(\boldsymbol{x}_t \mid c, t) := \nabla_{x_t} \log p_{\text{real},t}(\boldsymbol{x}_t \mid c), \qquad s_{\text{fake}}(\tilde{x}_t \mid c, t) := \nabla_{\tilde{\boldsymbol{x}}_t} \log p_{\text{fake},t}(\tilde{\boldsymbol{x}}_t \mid c), \tag{14}$$

and the teacher is a neural network $s_\phi(\cdot, c, t)$ approximating $s_{\text{real}}(\cdot \mid c, t)$.

The DFD and KL gradient at noise level t are:

$$\begin{aligned} g_{\text{DFD}}(\theta) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{z},\epsilon}\left[\big(s_{\text{fake}}(\tilde{\boldsymbol{x}}_t \mid c, t) - s_{\text{real}}(\boldsymbol{x}_t \mid c, t)\big)\nabla_\theta G_\theta(\boldsymbol{z}, c)\right], \\ \nabla_\theta D_{\text{KL}} &= \mathbb{E}_{\boldsymbol{z},\epsilon}\left[\big(s_{\text{fake}}(\tilde{\boldsymbol{x}}_t \mid c, t) - s_{\text{real}}(\tilde{\boldsymbol{x}}_t \mid c, t)\big)\nabla_\theta G_\theta(\boldsymbol{z}, c)\right]. \end{aligned} \tag{15}$$

(For brevity we have absorbed any α_t chain-rule factors into $\nabla_\theta G_\theta$ without changing the structure of the argument.)

CONDITION

Condition 1 (Bounded conditional matching error)**.** *There exists $\delta(c) \geq 0$ such that*

$$\mathbb{E}\Big[\|\boldsymbol{x} - G_\theta(\boldsymbol{z}, c)\|^2 \;\Big|\; c\Big] \leq \delta(c)^2. \tag{16}$$

Condition 2 (Lipschitz teacher score on the noised diffusion path)**.** *At each noise level $t > 0$, the teacher score network $s_{\text{real}}(\cdot \mid c, t)$ is $L(c, t)$-Lipschitz:*

$$\|s_{\text{real}}(\boldsymbol{u} \mid c, t) - s_{\text{real}}(\boldsymbol{v} \mid c, t)\| \leq L(c, t)\ \|\boldsymbol{u} - \boldsymbol{v}\|, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d. \tag{17}$$

This is justified for two complementary reasons: (i) the noised marginal $p_{\text{real},t}$ is the convolution of p_{real} with a Gaussian of variance σ_t^2, whose score has Hessian operator norm bounded by $\mathcal{O}(1/\sigma_t^2)$; (ii) the score is parameterized by a neural network with Lipschitz architectural primitives, so it inherits a finite Lipschitz constant on any bounded domain.

Condition 3 (Bounded generator Jacobian)**.** $\|\nabla_\theta G_\theta(\boldsymbol{z}, c)\|_{\text{op}} \leq B$ *almost surely.*

MAIN RESULT

Proposition 1. *Under Condition 1–3, for every noise level $t > 0$,*

$$\left\|g_{\mathrm{DFD}}(\theta) - \nabla_\theta D_{\mathrm{KL}}(p_{\mathrm{fake}} \| p_{\mathrm{real}})\right\|_2^2 \;\le\; B^2\, \alpha_t^2\, L(c,t)^2\, \delta(c)^2. \tag{18}$$

Proof. **Step 1: Cancellation of the fake score.** The fake-score term in (**??**) depends only on $(\boldsymbol{z}, \epsilon, c)$, not on $\boldsymbol{x}$, so $\mathbb{E}_{\boldsymbol{x},\boldsymbol{z},\epsilon|c}[\,\cdot\,] = \mathbb{E}_{\boldsymbol{z},\epsilon|c}[\,\cdot\,]$ for that term. It cancels with the corresponding term in (**??**), giving

$$g_{\mathrm{DFD}}(\theta) - \nabla_\theta D_{\mathrm{KL}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z},\epsilon|c}\big[\big(s_{\mathrm{real}}(\tilde{\boldsymbol{x}}_t \mid c, t) - s_{\mathrm{real}}(\boldsymbol{x}_t \mid c, t)\big)\nabla_\theta G_\theta(\boldsymbol{z}, c)\big]\,. \tag{19}$$

Step 2: Jensen's inequality. Since $\|\cdot\|_2^2$ is convex,

$$\left\|g_{\mathrm{DFD}}(\theta) - \nabla_\theta D_{\mathrm{KL}}\right\|_2^2 \;\le\; \mathbb{E}_{\boldsymbol{x},\boldsymbol{z},\epsilon|c}\Big[\big\|\big(s_{\mathrm{real}}(\tilde{\boldsymbol{x}}_t \mid c, t) - s_{\mathrm{real}}(\boldsymbol{x}_t \mid c, t)\big)\nabla_\theta G_\theta(\boldsymbol{z}, c)\big\|_2^2\Big]\,. \tag{20}$$

Step 3: Bounding the integrand. Using sub-multiplicativity, Condition 2, and Condition 3,

$$\begin{aligned}\big\|\big(s_{\mathrm{real}}(\tilde{\boldsymbol{x}}_t \mid c, t) - s_{\mathrm{real}}(\boldsymbol{x}_t \mid c, t)\big)\nabla_\theta G_\theta(\boldsymbol{z}, c)\big\|_2^2 &\le B^2\, \|s_{\mathrm{real}}(\tilde{\boldsymbol{x}}_t \mid c, t) - s_{\mathrm{real}}(\boldsymbol{x}_t \mid c, t)\|^2 \\ &\le B^2\, L(c,t)^2\, \|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\|^2\,.\end{aligned} \tag{21}$$

Since $\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t = \alpha_t\big(G_\theta(\boldsymbol{z}, c) - \boldsymbol{x}\big)$ (the noise term $\sigma_t\epsilon$ is shared and cancels),

$$\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\|^2 = \alpha_t^2\, \|G_\theta(\boldsymbol{z}, c) - \boldsymbol{x}\|^2\,. \tag{22}$$

Step 4: Apply the matching bound. Combining the previous steps,

$$\begin{aligned}\left\|g_{\mathrm{DFD}}(\theta) - \nabla_\theta D_{\mathrm{KL}}\right\|_2^2 &\le B^2\, \alpha_t^2\, L(c,t)^2\, \mathbb{E}\Big[\|\boldsymbol{x} - G_\theta(\boldsymbol{z}, c)\|^2 \;\big|\, c\Big] \\ &\le B^2\, \alpha_t^2\, L(c,t)^2\, \delta(c)^2. \quad \square\end{aligned} \tag{23}$$

which is equavalent to:

$$\mathbb{E}\big[\big\|\Delta_{p_{\mathrm{real}},\boldsymbol{x},G_\theta(\boldsymbol{z})}\big\|_2^2 \,\big|\, c\big] \;\lesssim\; \alpha_t^2\, L(c,t)^2\, \delta(c)^2 \tag{24}$$

B. Experiment Details

B.1. Data Curation

We mainly curated two datasets for our experiments:

- *animation dataset*: video clips with an animation or cartoon visual style.
- *mix-style dataset*: video clips spanning all visual styles.

Both are derived from the ViPE dataset (Huang et al., 2025a), starting from ∼966k captioned source clips. Our pipeline has five stages: (1) tri-level captioning, (2) CLIP visual feature extraction, (3) K-means clustering, (4) per-cluster grid visualization with human-in-the-loop selection, and (5) WebDataset packaging at three caption granularities. The full pipeline code is released alongside the paper.

B.1.1. TRI-LEVEL CAPTIONING.

For each source video we generate three captions of different lengths in a single VLM forward pass, using **Qwen3-VL-4B-Instruct** served via vLLM with continuous batching and PagedAttention. A single system prompt instructs the model to emit

- `[LONG]`: a 250–300 word caption covering subject appearance, actions, expressions, environment, camera work, and visual style;
- `[MEDIUM]`: a 150–180 word caption that retains the key subject, actions, setting, and style;
- `[SHORT]`: a single sentence capturing the essence of the clip.

Captions are decoded with sampling parameters $T = 0.7$, top-$p = 0.8$, and `max_new_tokens`$= 512$. Producing all three lengths in one forward pass (rather than three independent passes) reduces VLM cost roughly $3\times$ since the visual tokens are encoded only once. The three captions are parsed from the tagged output via regular expressions; if any tag is missing we fall back deterministically (truncate the long caption for medium, take the first sentence for short). Each caption is also encoded with the UMT5-XXL text encoder (Wan2.1-T2V-1.3B checkpoint) and the resulting prompt embeddings are stored alongside the raw text.

B.1.2. Visual Feature Extraction.

For diversity sampling we want a feature space that reflects visual content rather than caption phrasing. We therefore extract CLIP ViT-B/32 image feature. For each video we decode 3 uniformly-spaced keyframes with PyAV, apply the standard CLIP preprocessing (`Resize`(224), `CenterCrop`(224), `ToTensor`, CLIP normalization), encode each keyframe with `encode_image`, mean-pool the three frame embeddings, and ℓ_2-normalize, yielding one 512-d vector per video.

B.1.3. Clustering.

We cluster the $\sim$966k feature vectors with **FAISS** K**-means** (K=1000, 50 iterations). Because all vectors are ℓ_2-normalized, squared Euclidean distance to the centroid is monotonically related to cosine similarity, so we use it directly for ranking inside each cluster. The output is a per-video integer cluster label and a $(K,512)$ centroid matrix.

B.1.4. Per-Cluster Visualization and Human-in-the-Loop Selection.

For each of the K=1000 clusters we (i) sort all members by Euclidean distance from the centroid, (ii) pick 40 candidates at evenly-spaced percentiles of that distance (so row 0 is the centroid-nearest video and row 39 is the farthest), and (iii) decode 5 evenly-spaced frames per candidate. A human reviewer then opens each cluster grid and writes the row indices of 20 visually-representative-but-distinct videos into the `picks` column, e.g. `"0,2,5,7,9,12,...,38,39"`. The even-percentile sampling ensures the 40 candidates already span the full intra-cluster spread, so the reviewer is choosing among visually distinct options rather than near-duplicates. Clusters whose grid contains no useful content (e.g., uniformly black frames, decode-failed clusters) can be left empty and are dropped from the final dataset; this trades a small number of slots for higher quality. After review we obtain 30,000 selected videos in cluster-id order.

The animation dataset is constructed by the same filtering pipeline, restricted to clusters whose grid visualizations are dominated by animation/cartoon style during the human review step.

B.2. Baseline Details

We show the implementation details of our baselines in Table 4. For the text-to-video tasks, all the baseline and our method are trained on 8 A100 GPUs. For the image-to-video tasks, all the baseline and our method are trained on 16 A100 GPUs.

B.3. Text-to-Video

Metric Details We show the metric details here. For each text prompt we sample $L = 8$ videos from distinct initial noise seeds, and report three families of metrics: a visual diversity metric adapted from DP-DMD, a set of VBench (Huang et al., 2024) quality dimensions, and camera-pose statistics extracted with ViPE (Huang et al., 2025a).

B.3.1. Visual Diversity Metric.

We adopt the diversity metric from DP-DMD (Wu et al., 2026). We assess sample diversity using DINOv2-ViT-Large (DINO) (Siméoni et al., 2025) and CLIP-ViT-Large (CLIP) (Radford et al., 2021) by computing the cosine similarity

Table 4. Training configurations.

Models	Cosmos Predict2 I2V			Wan2.1 T2V		
	DMD2	**DP-DMD**	**DFD**	**DMD2**	**DP-DMD**	**DFD**
Num. of Frame	81	81	81	81	81	81
Batch Size	16	16	16	16	16	16
Resolution	480p	480p	480p	480p	480p	480p
Learning Rate(discriminator)	1e-05	1e-05	N/A	1e-05	1e-05	N/A
Learning Rate (student)	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05
Learning Rate (fake score)	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05
CFG Scale	3.0	3.0	3.0	5.0	5.0	5.0
Student Update Frequency	5	5	5	5	5	5
Diversity anchor step	N/A	1	N/A	N/A	5	N/A
Diversity weights	N/A	0.05	N/A	N/A	0.05	N/A
Total Iterations	30k	30k	300	30k	30k	100
Pretrained Model Iterations	0	0	25k(from DMD2)	0	0	25k(from DMD2)

between extracted image feature representations in a pairwise manner:

$$\text{Diversity} = 1 - \frac{2}{L(L-1)} \sum_{i,\,j} \cos\left(\boldsymbol{x}_\theta^{(i)}, \boldsymbol{x}_\theta^{(j)}\right), \tag{25}$$

where L denotes the number of distinct initial noise samples per text prompt[1], and we set $L = 8$ in our experiments. To adapt this image-level metric to videos, we sample 16 evenly-spaced frames from each generated clip, extract per-frame CLIP and DINO embeddings, and mean-pool them into a single feature vector $\boldsymbol{x}_\theta^{(i)}$ per video before applying Eq. 25. The reported diversity is the average over all text prompts.

B.3.2. VBench Quality Metrics.

For t2v tasks, We evaluate generation quality with VBench (Huang et al., 2024) which scores videos directly without requiring access to the original training prompts. We report the seven prompt-free dimensions:

- *subject consistency*: measures whether the main subject's appearance (identity, shape, texture) remains stable across frames, computed via DINO feature similarity between frames.
- *background consistency*: measures temporal stability of the background scene across frames using CLIP feature similarity, capturing flicker or drift in the surrounding environment.
- *temporal flickering*: penalizes high-frequency, low-level fluctuations between adjacent frames in static or near-static regions, reflecting visible flicker artifacts.
- *motion smoothness*: assesses whether motion follows the priors of a video frame interpolation model, with smoother (more physically plausible) motion scoring higher.
- *dynamic degree*: estimates the amount of motion in the video using optical flow magnitude, rewarding generations that are not nearly-static.
- *aesthetic quality*: predicts the human-perceived aesthetic score of sampled frames using the LAION aesthetic predictor.
- *imaging quality*: predicts low-level image quality (sharpness, noise, distortion) using the MUSIQ image-quality predictor.

For i2v tasks, we evaluate the following metrics provided in VBench i2v part:

[1]For each text prompt, every pair of generated samples is compared once, yielding $\binom{L}{2}$ total comparisons.

- *subject consistency:* Measures the model's ability to preserve the identity, appearance, and structural integrity of the primary subject across all frames of the generated video.
- *Background Consistency:* Evaluates the temporal stability and visual coherence of the background environment, penalizing unnatural shifts or morphing over time.
- *Aesthetic Quality:* Assesses the overall visual appeal, lighting, composition, and artistic fidelity of the individual generated frames.
- *Temporal Flickering:* Quantifies the presence of high-frequency visual artifacts or unnatural, rapid changes in texture and lighting between adjacent frames.
- *Motion Smoothness:* Evaluates the naturalness, fluidity, and physical plausibility of the movements occurring within the video sequence.
- *Image-to-Video Subject:* Specifically for Image-to-Video (I2V) generation tasks, this metric calculates how accurately the subject in the generated video aligns with the reference subject provided in the conditioning input image.
- *Image-to-Video Background:* Evaluates how well the generated video maintains and reflects the background elements present in the initial conditioning image.

B.3.3. CAMERA-POSE DIVERSITY.

To quantify how the camera moves across generated videos, we run ViPE (Huang et al., 2025a) on every clip to recover an 81-frame sequence of 4×4 camera-to-world transformation matrices $\{T_t\}_{t=0}^{80}$, with translation $\boldsymbol{p}_t = T_t[:3, 3]$ and rotation $R_t = T_t[:3, :3]$. We then derive two groups of metrics from these poses.

Per-video motion magnitude (computed independently for each video):

- *translation path length*: $\sum_{t=0}^{N-2} \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|$, the total distance the camera travels.
- *max translation*: $\max_t \|\boldsymbol{p}_t - \boldsymbol{p}_0\|$, the farthest the camera reaches from its starting position.
- *rotation total* (degrees): $\sum_{t=0}^{N-2} \angle\big(R_{t+1}R_t^\top\big)$, the accumulated frame-to-frame rotation magnitude.
- *max rotation* (degrees): $\max_t \angle\big(R_tR_0^\top\big)$, the largest deviation from the initial orientation.

Cross-seed trajectory diversity (computed per prompt across the $L = 8$ seeds, then averaged over prompts). Let $\boldsymbol{e}^{(i)} = \boldsymbol{p}_{N-1}^{(i)}$ be the endpoint of the i-th seed and $\{\boldsymbol{p}_t^{(i)}\}$ its full trajectory:

- *endpoint spread*: $\big\|\mathrm{std}_i \boldsymbol{e}^{(i)}\big\|_2$, the ℓ_2 norm of the per-axis standard deviation of the endpoints.
- *mean pairwise endpoint distance*: $\frac{2}{L(L-1)} \sum_{i<j} \|\boldsymbol{e}^{(i)} - \boldsymbol{e}^{(j)}\|$, averaged over all $\binom{L}{2} = 28$ seed pairs.
- *mean pairwise trajectory distance*: $\frac{2}{L(L-1)} \sum_{i<j} \frac{1}{N} \sum_t \|\boldsymbol{p}_t^{(i)} - \boldsymbol{p}_t^{(j)}\|$, the average per-frame distance between trajectories, capturing path-level (not just endpoint) differences.

C. Additional Results

C.1. Additional Results for Text-to-Video Generation

We show more visualization results for the text-to-video generation. And **We refer the readers to the supplementary material for full video comparisons**

C.1.1. DIVERSITY VISUALIZATION RESULTS

We further illustrate the diversity advantage of our method by generating videos with 8 different random seeds under the same text prompt. Fig. 10 shows the middle frame of each generated video for DMD2, DP-DMD, and ours. Each row corresponds to a different pair of seeds (rows 1–4 cover seeds 1–8). Within each method, the two columns show two different seeds; within each row, the same two seeds are shown across the three methods. Compared to DMD2 and DP-DMD, our method produces clearly more diverse outputs across seeds, with larger variation in subject pose, layout, and scene composition.



Figure 10. **Diversity visualization across 8 random seeds for the same prompt**. Columns 1–2: DMD2; columns 3–4: DP-DMD; columns 5–6: Ours. Each row shows the middle frame of two videos generated with different seeds. Our method produces visibly more diverse outputs across seeds.

C.2. Main Results for Auto

C.3. Additional Results for Image-to-Video Generation

Table 5. Quantitative comparison on image-to-video generation. We evaluate the metrics from VBench on our curated ViPE test set. For all metrics, higher is better. The best result in each column is **bolded**. Our method consistently outperforms DMD2 and DP-DMD, and is even better than the teacher model.

Method	Subject Consistency	Background Consistency	Aesthetic Quality	Temporal Flickering	Motion Smoothness	Image-to-Video Subject	Image-to-Video Background	Average
Teacher	0.9340	0.9468	0.6026	0.9694	0.9886	0.9781	0.9860	0.9151
DMD2	0.9340	0.9473	**0.6095**	0.9722	0.9892	0.9805	0.9868	0.9171
DP-DMD	0.8372	0.8942	0.5621	0.9562	0.9823	0.9545	0.9655	0.8789
Ours	**0.9417**	**0.9538**	0.6047	**0.9814**	**0.9918**	**0.9825**	**0.9893**	**0.9207**

We refer the readers to the supplementary material for full video comparisons

C.4. Additional Results for ablation study

Table 6. Ablation on the effects of GAN loss in our DFD. Bold indicates the better value. Removing the GAN loss yields results comparable to the model distilled with the GAN loss.

Model (DFD)	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
w/o GAN	**0.9690**	0.9620	0.9785	0.9899	**0.5000**	0.7194	**0.7452**
w GAN	0.9666	**0.9625**	**0.9831**	**0.9912**	0.3750	**0.7213**	0.7210

Table 7. Ablation on the weights in Eq. 9. The results show no clear difference between the two choices of w.

Model	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
$w = \frac{1}{2}$	**0.9691**	**0.9661**	0.9830	0.9907	**0.5625**	0.7135	**0.7457**
$w = 1$	0.9666	0.9625	**0.9831**	**0.9912**	0.3750	**0.7213**	0.7210

C.4.1. TEXT PROMPTS

We show part of the text prompts here:

Example Prompts

- Pikachu and Eevee dancing joyfully together on a dirt path in a bright, sunlit forest.
- Cheerful yellow cartoon dragon with green spikes and colorful wings, standing on a bright grassy hill.
- SpongeBob SquarePants dancing energetically inside his pineapple house, surrounded by floating bubbles.
- Cartoon-style family of polar bears playing on a floating ice floe.
- Animated squirrel gathering acorns in a colorful autumn forest.
- Anime-style mischievous cat batting at a ball of yarn in a cozy living room.
- Animation of a graceful fox running through a snowy landscape.
- Cartoon-style happy dog playing fetch on a bright sandy beach.
- Animated traditional Japanese village at sunset with glowing paper lanterns and falling cherry blossoms.
- Animated misty village with green-roofed wooden houses, a cobblestone path, and vibrant orange flowers.
- An animated princess and men in tuxedos interacting cheerfully by a grand clock tower at night.
- Cartoon-style elaborate underwater palace made of shells and coral.
- Animated complex European castle perched on a rugged, mist-shrouded mountain peak.
- Anime-style bustling Asian street market with colorful awnings, glowing signs, and dense pedestrian traffic.
- Cartoon-style quirky, multi-level treehouse with suspension bridges and tire swings in a forest.
- Animation of a historic lighthouse on a rocky coast battling powerful stormy waves.
- Anime-style Sailor Moon walking down a neon city street, looking back over her shoulder with a smile.
- Cheerful snowman standing on a snowy roof, looking down at a lit Christmas tree in a town square.
- Anime close-up of an angry girl with pink twin-tails, golden horns, and a dark purple outfit.
- Anime scene of a grinning Naruto holding ramen, standing next to a smiling girl with lavender hair.
- Cartoon blonde mermaid with a white tiara, swimming happily through a vibrant coral reef.
- Three animated women excitedly examining a map at a table in an ornate room.
- A cartoon woman smiling and sipping coffee from a red cup at a sunny, cozy café.
- A romantic animated couple gazing at each other under a Van Gogh-style swirling starry night sky.
- Two animated princesses having a magical outdoor tea party among lush flowers and glowing sparkles.
- Two animated princesses dancing joyfully on a palace terrace in the rain under a vibrant rainbow.
- An animated princess pointing enthusiastically toward a seated queen at a magical royal banquet.
- An animated princess pointing enthusiastically toward a seated queen at a magical royal banquet.
- Animation of a street artist creating a detailed vibrant mural on a brick wall.

- An animated woman in a blue gown gazing from a rose-adorned balcony overlooking a fairy-tale city.
- A cheerful cartoon princess gesturing joyfully in a vibrant, sunny garden with a distant palace.
- A tearful cartoon princess in a starry gown looking up at the night sky, with a majestic castle behind her.
- Cartoon-style towering waterfall plunging into a clear turquoise pool surrounded by jungle foliage.
- Animation of a vast desert with towering sand dunes stretching to the horizon under a blazing sun.
- Anime-style dense, ancient forest where beams of sunlight filter dramatically through the high canopy.
- Cartoon-style bustling cityscape viewed from a high rooftop, with lights blinking on as twilight descends.
- Animation of a tranquil lake perfectly reflecting the purple and orange colors of a vibrant dawn.
- Cartoon pig chef wearing a tall white hat, proudly holding a strawberry cake in a modern kitchen.
- Cartoon-style towering stack of pancakes dripping with maple syrup and melting butter in a cozy diner.
- Animation of a detailed chef's market stall with rows of colorful fresh fruits and vegetables.
- Anime-style detailed close-up of a steaming hot bento box full of intricate, colorful food.
-
- Cartoon-style conveyor belt with tiny, perfectly formed sushi plates moving past happy diners.
- Animated bakery window display filled with intricate pastries, artisanal bread, and glowing warmth.
- Cartoon-style colorful ice cream truck with children lining up on a hot summer day.
- Anime-style traditional tea ceremony performed in a peaceful garden with gentle steam rising from the cups.
- Cartoon-style bright red vintage steam train chugging across rolling green hills under a bright sky.
- Animated futuristic spaceship smoothly landing on a alien planet with purple vegetation.
- Whimsical stop-motion style yellow submarine exploring deep, glowing underwater ruins.
- Anime depiction of a sleek sports car speeding through a tunnel illuminated by passing neon lights.
- Cartoon-style biplane flying dynamic loop-the-loops amidst fluffy, stylized white clouds.
- Animated pirate ship with detailed sails navigating a vast, sun-drenched ocean at sunset.
- Cartoon delivery scooter expertly navigating a bustling, rainy city street.
- Anime-style flying motorcycle soaring high above a dense, green forest canopy.
- Cartoon-style dusty antique clock shop with dozens of clocks ticking and chiming simultaneously.
- Animated messy artist's desk covered in used paintbrushes, tubes of paint, and open sketchbooks.
- Whimsical stop-motion style pair of worn leather boots resting by a crackling fireplace.
-
- Anime-style close-up of a well-traveled backpack adorned with colorful keychains and buttons.
- Cartoon-style bookshelf packed with a vast, eclectic collection of books in a library corner.
- Animation of a grand grandfather clock with its large pendulum slowly swinging back and forth.

- Cartoon-style gardening tools—a trowel, gloves, and seed packets—resting on a rustic wooden bench.
- Anime-style glowing crystal globe positioned on a wooden desk within a wizard's library.
- Animated messy child's playroom filled with various toys and building blocks scattered on the floor.
- Anime-style depiction of a massive magical library where books fly autonomously between high shelves.
- Animation of a whimsical city made of clouds floating in the sky, connected by glowing bridges of light.
- Cartoon-style encounter with a friendly forest spirit in a glowing, ancient grove of mossy trees.
- Anime-style powerful wizard casting a complex spell that generates swirling vortexes of magical energy.
- Animation of a mischievous imp hiding behind colorful, glowing mushrooms within a deep cave.
- Cartoon-style friendly dragon and a knight playing chess peacefully under the shade of a large tree.
- Anime-style phoenix rising dramatically from a bed of glowing ashes with brilliant, fiery feathers.
- Animation of an intricate city built entirely from colorful crystals that hum and pulse with light and energy.
- Whimsical cartoon depiction of a celestial garden where stars grow like flowers on glowing vines.
- Whimsical cartoon depiction of a celestial garden where stars grow like flowers on glowing vines.
- A young woman with curly red hair laughs while spinning a yellow umbrella in slow motion as rain falls around her on a cobblestone street.
- Close-up of an elderly man's weathered hands carefully tying a fly fishing lure beside a misty mountain stream at dawn.
- Cinematic shot of a lone figure in a long coat walking across an endless salt flat under a vast purple twilight sky.
- A barista pulls a perfect espresso shot, dark crema swirling into a white ceramic cup in soft morning light.
- Drone footage soaring above a turquoise coastline where waves crash against jagged white cliffs.
- A chef in a black apron flips a sizzling steak in a cast iron pan, flames leaping briefly upward.
- Close-up of a honeybee landing on a sunflower, pollen visibly clinging to its legs as it crawls across the petals.
- A vintage motorcycle rolls slowly down an empty desert highway at sunset, heat shimmering off the asphalt.
- Cinematic tracking shot of a ballerina in a white tutu rehearsing alone in a sunlit wooden studio.
- Time-lapse of storm clouds rolling rapidly across a wheat field as the wind bends the golden stalks.
- A black cat slinks silently across a wet rooftop at night, eyes glowing under a streetlamp.
- Macro shot of a single dewdrop sliding down a green blade of grass at sunrise.
- A woman in a flowing red dress runs barefoot along a deserted beach, footprints quickly washed away by the surf.
- First-person view from inside a glass elevator rising rapidly up a futuristic skyscraper at night.
- A young boy lights sparklers in his backyard as twilight settles, his face illuminated in flickering gold.

- Aerial shot of a small wooden boat carving a wake across a glassy lake surrounded by dense pine forest.
- A potter's hands shape wet clay on a spinning wheel, water and clay slipping between their fingers.
- Cinematic close-up of an astronaut's helmet visor reflecting the curve of Earth from low orbit.
- A street musician plays a worn acoustic guitar on a busy sidewalk as commuters hurry past in motion blur.
- Slow-motion shot of a basketball swishing through a chain net on an outdoor court at golden hour.
- A fox steps cautiously into a snowy clearing at dusk, ears twitching as snowflakes drift down around it.
- Wide shot of a Tibetan monastery perched on a cliff edge, prayer flags whipping in the high mountain wind.
- A barista pours latte art into a cappuccino, viewed from directly above as a delicate rosetta forms.
- A child runs through a sprinkler in a suburban backyard on a hot summer afternoon, laughing.
- Macro footage of ink spreading and blooming through clear water, forming twisting black tendrils.
- Cinematic shot of a saxophonist playing on a foggy New Orleans street corner, neon signs glowing behind him.
- An elderly couple slow dances in their kitchen as warm afternoon light streams through gauzy curtains.
- Drone shot pulling back from a single red farmhouse to reveal endless rolling vineyards stretching to the horizon.
- A blacksmith hammers a glowing orange blade on an anvil, sparks flying with each strike.
- Underwater shot of a sea turtle gliding gracefully through shafts of sunlight piercing a coral reef.
- A businesswoman in a tailored suit strides confidently across a rain-soaked Tokyo crosswalk at night, neon reflecting in puddles.
- Close-up of a vinyl record dropping onto a turntable, the needle settling as it begins to spin.
- Time-lapse of a city skyline transitioning from sunset to night, windows lighting up across the buildings.
- A surfer paddles into a massive wave at dawn, silhouetted against a pink and orange sky.
- Cinematic shot of a horse galloping across a misty Scottish moor, hooves throwing up clumps of damp earth.
- A jazz pianist's hands move expressively across the keys in a dimly lit speakeasy filled with soft amber light.
- A baker dusts flour across a freshly braided loaf of bread on a wooden countertop in a sunlit kitchen.
- Slow-motion shot of a tennis ball striking a racket, the strings flexing dramatically on impact.
- A father teaches his young daughter to ride a bicycle along a tree-lined neighborhood street in autumn.
- Aerial view of a fishing village at dawn, colorful boats returning to harbor across calm waters.
- Macro shot of a butterfly slowly opening and closing its wings on a lavender stem in a summer breeze.
- A man reads a leather-bound book in a worn armchair beside a crackling fireplace as snow falls past the window.
- Cinematic shot of a train emerging from a tunnel into a sunlit mountain valley filled with wildflowers.

- A surgeon in scrubs washes their hands at a hospital sink, water cascading over their fingers in clinical white light.
- Close-up of raindrops striking a car windshield, the wipers sweeping rhythmically across the glass.
- A florist arranges a bouquet of peonies and eucalyptus on a marble counter in a bright minimalist shop.
- A lone hiker reaches a mountain summit at sunrise and stands silhouetted against the rising sun, arms outstretched.
- Slow-motion shot of a glass shattering as it strikes a tile floor, shards spreading outward.
- Drone footage following a red kayak winding through a narrow emerald-green river canyon.
- A young man in a leather jacket leans against a vintage car at a 1950s-style diner, neon signs buzzing overhead.
- Cinematic shot of two dancers performing a tango in an empty warehouse lit by a single overhead bulb.
- A child's hand reaches up to touch the glass of an aquarium as a stingray glides past on the other side.
- Wide shot of a hot air balloon festival at dawn, dozens of colorful balloons inflating across a misty field.
- A glassblower shapes molten glass at the end of a long iron rod, the orange glow lighting their focused face.
- Close-up of espresso dripping into a clear glass mug, steam curling upward in soft window light.
- A street artist spray-paints a vibrant mural on a brick wall in an alley, bystanders watching in the background.
- Slow-motion shot of a diver leaping from a high cliff into a deep blue ocean below.
- A woman pours steaming tea from a cast iron kettle into a delicate porcelain cup at a wooden table.
- Aerial shot of an autumn forest in full color, a winding road cutting through a fiery red and orange canopy.
- A construction worker welds a steel beam high above a city skyline, sparks raining down against the sunset.
- Cinematic shot of a vintage typewriter on a desk by a rainy window, paper slowly being pulled from the carriage.
- A skateboarder grinds along a metal handrail in an urban plaza, captured in slow motion at golden hour.
- A grandmother teaches her granddaughter to roll dumplings at a flour-dusted kitchen counter, both smiling.
- Underwater shot of a school of silver fish parting around a diver swimming through a sunlit kelp forest.
- Time-lapse of a flower blooming from a tight bud to a full open blossom over the course of a day.
- A scientist examines a glowing blue liquid in a test tube inside a softly lit laboratory.
- Cinematic wide shot of a desert caravan crossing massive sand dunes at sunset, long shadows trailing behind.
- A coffee shop window steams up from the inside as rain streaks down the outside, a person reading at the counter.
- Slow-motion shot of a flock of birds taking off from a marsh at dawn, wings beating against pink mist.
- A young woman pins a Polaroid photo onto a string of lights in her bedroom, the photo gently swinging.
- Drone footage gliding low over a frozen lake at dusk, a single ice skater carving graceful arcs across its surface.

C.5. Societal Impacts

Our work on distilling video diffusion models presents both positive and negative potential societal impacts. On the positive side, our method significantly reduces the computational cost and inference time required for high-fidelity video generation. This democratizes access to creative tools, allowing users with limited compute resources to generate videos, while also reducing the carbon footprint and energy consumption associated with deploying large-scale generative models. On the negative side, accelerating video generation inherently scales the risks associated with the base models. Faster generation can facilitate the rapid creation of deepfakes, misinformation, and malicious content. While our theoretical framework improves algorithmic efficiency, it does not inherently prevent these misuses. Safe deployment of these distilled models in real-world applications will require coupling them with robust safety filters, provenance tracking, and watermarking techniques.