RETHINKING SELF-FORCING DMD FROM A SCORE SHARPNESS PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-Forcing with Distribution Matching Distillation (DMD) enables real-time autoregressive video generation, but its outputs can exhibit insufficient motion and degraded visual quality. We revisit these limitations from a *score sharpness* perspective and use this perspective to motivate changes to the DMD training signal. An illustrative one-dimensional Gaussian mixture connects the noise level to local variations of the score, suggesting a timestep curriculum that gradually shifts training from high-noise to low-noise regimes. We further introduce a timestep offset with antithetic sampling, motivated by the hypothesis that small changes to the teacher's input timestep may mitigate some prediction errors, and adopt the projection mechanism of *Adaptive Projected Guidance* (APG) to down-weight the parallel guidance component. These three interventions act during training and leave the student architecture and inference procedure unchanged. Experiments on Wan2.1-T2V-1.3B show higher Total, Quality, and Semantic scores than Self-Forcing under the reported evaluation protocol, while preserving real-time throughput. Our method obtains the highest reported Total and Quality scores among the compared methods, with a lower Semantic score than Causal Forcing. Component ablations support the benefits of combining the interventions, including an increase in Dynamic Degree from 23.12 to 56.48 on the ablation prompt suite.

1 INTRODUCTION

Diffusion-based video generation has made remarkable progress (Wan et al., 2025; HaCohen et al., 2024; Deng et al., 2025; Jin et al., 2025; Chen et al., 2025; Teng et al., 2025), yet the standard bidirectional denoising paradigm requires iterating over all frames simultaneously, leading to high latency and memory cost. Autoregressive video generation offers a scalable alternative by producing frames causally via a KV-cache, supporting streaming synthesis without jointly denoising an entire video.

Self-Forcing (Huang et al., 2026) reduces the train-test mismatch by training on the student's own outputs and combines this with Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) to achieve real-time generation. However, its outputs can exhibit insufficient motion and visual degradation, including texture artifacts and loss of fine detail. These symptoms motivate a closer examination of the training signal used during distillation. Mode-seeking behavior provides a possible interpretation of limited variation. We focus on improving motion and visual quality through changes to the distillation signal.

We revisit Self-Forcing DMD from a *score sharpness* perspective. We use sharpness to describe how rapidly the score changes with its input, rather than treating score magnitude as a direct measure of sharpness. In an illustrative one-dimensional Gaussian mixture, the score varies more sharply between modes at low noise levels than at high noise levels. This observation motivates changing when the student encounters different noise levels. We also consider two teacher-side factors: prediction errors that may change when the teacher's input timestep is slightly increased or decreased, and the parallel guidance component that may amplify the teacher signal under classifier-free guidance (CFG) (Ho & Salimans, 2022). Together, these considerations motivate three training interventions.

Guided by this perspective, we develop a **timestep curriculum** that evolves a Beta sampling distribution from high-noise to low-noise regimes and a **timestep offset** that applies positive and negative

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

perturbations to the teacher's input timestep. We additionally adopt the projection mechanism of **Adaptive Projected Guidance** (APG) (Sadat et al., 2024), down-weighting the guidance component parallel to the conditional score while retaining its orthogonal component. Together, these interventions modify timestep sampling, teacher evaluation, and guidance composition during DMD training, without changing the student architecture or inference procedure. Our contributions are as follows:

- We develop a score sharpness perspective on the training signal in Self-Forcing DMD, using a one-dimensional analysis to motivate hypotheses about motion and visual quality degradation.
- We introduce a timestep curriculum and a teacher timestep offset, and integrate an APG-based guidance modification into DMD training. The resulting method requires no changes to the inference pipeline.
- Experiments show improved motion and visual quality metrics under the reported settings. Our method obtains the highest Total and Quality scores among the compared methods, while component ablations support combining the three interventions and guidance comparisons characterize their trade-offs.

2 Method: Rethinking the DMD Training Signal

2.1 Preliminaries

Self-Forcing (Huang et al., 2026) is a training paradigm for causal video generation in which the student generator G_θ autoregressively produces latent frames $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ from noise, conditioning each frame on its own previously generated outputs via a causal KV-cache mechanism. This closes the train-test gap inherent in teacher-forcing approaches.

Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) distills the student by minimizing the KL divergence between the student distribution p_θ and the data distribution p_{data} through score matching. Given a generated sample $\mathbf{x} \sim p_\theta$, a random re-noising timestep t is sampled and noise is added to obtain $\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$. The DMD gradient is:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} \propto \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\left(s_\phi(\mathbf{x}_t, t) - s_\psi(\mathbf{x}_t, t) \right) \frac{\partial \mathbf{x}}{\partial \theta} \right], \tag{1}$$

where s_ψ is the frozen teacher (real) score and s_ϕ is a learnable fake score trained on student-generated samples. In practice, the teacher score is enhanced with classifier-free guidance (CFG) (Ho & Salimans, 2022):

$$\tilde{s}_\psi(\mathbf{x}_t, t, \mathbf{c}) = s_\psi(\mathbf{x}_t, t, \mathbf{c}) + w \cdot \big(s_\psi(\mathbf{x}_t, t, \mathbf{c}) - s_\psi(\mathbf{x}_t, t, \varnothing) \big), \tag{2}$$

where $\mathbf{c}$ is the text condition, $\varnothing$ denotes the null condition, and w is the guidance scale.

When Self-Forcing is combined with DMD for video generation, its outputs can exhibit limited motion and degraded textures or fine details. We investigate training interventions motivated by a *score sharpness* perspective. In the rest of this section, we describe the motivating hypotheses (Sec. 2.2), provide an overview of our framework (Sec. 2.3), and detail each component (Secs. 2.4∼2.6).

2.2 A Score Sharpness Perspective on the Training Signal

In Self-Forcing DMD with CFG-enhanced teacher scores, the gradient that drives the student update at training step k takes the form:

$$\nabla_\theta \mathcal{L} \propto \mathbb{E}_{t \sim p(t|k),\, \boldsymbol{\epsilon}} \Big[\underbrace{\left(s_\phi(\mathbf{x}_t, t) - \tilde{s}_\psi(\mathbf{x}_t, t) \right)}_{\Delta(\mathbf{x}_t, t)} \frac{\partial \mathbf{x}}{\partial \theta} \Big], \tag{3}$$

where $p(t|k)$ is the timestep sampling distribution and $\tilde{s}_\psi$ is the guided teacher score. Both the fake and teacher scores contribute to Δ, and the resulting parameter update also depends on the generator Jacobian. We use *score sharpness* to refer to the local sensitivity of a score field to changes in its input. This differs from its magnitude at a point and from the concentration of update energy across samples or timesteps. These properties may interact during optimization, but one does not by itself

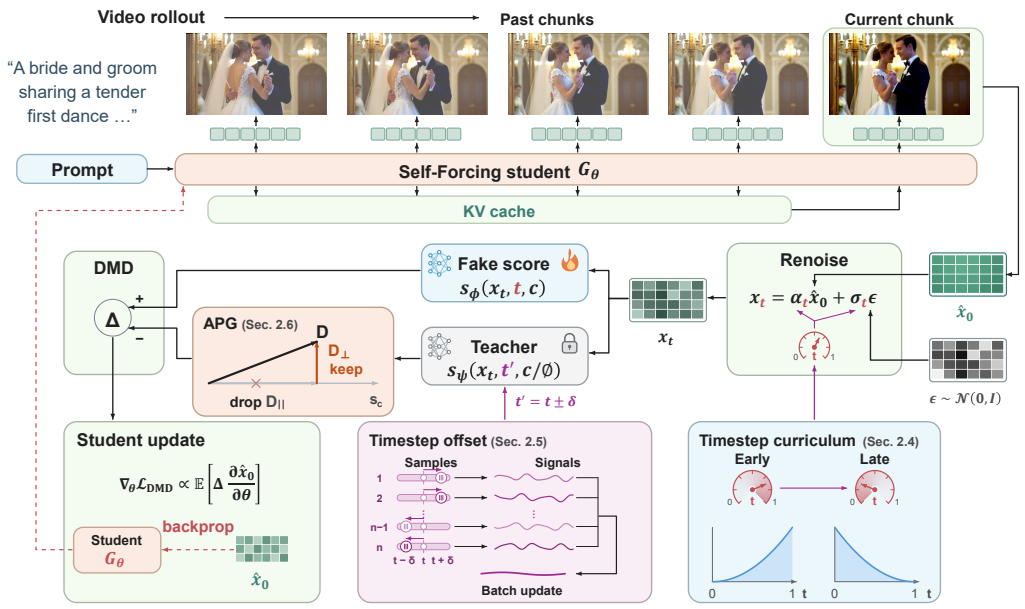


Figure 1: Overview of our framework. Three training-only interventions modify the DMD signal: a **timestep curriculum** reshapes the sampling distribution of t (Sec. 2.4); a **timestep offset** perturbs the teacher input timestep (Sec. 2.5); and **APG** down-weights the parallel guidance component (Sec. 2.6).

establish another. With this distinction, we consider three aspects of the training signal that motivate the interventions in Sections 2.4∼2.6.

(i) Noise level and score sharpness. Modern video models including the one used in Self-Forcing are trained under the rectified flow framework (Liu et al., 2022; Lipman et al., 2022), where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and $t \in [0, 1]$. The marginal is $p_t(\mathbf{x}) = \int p_{\text{data}}(\mathbf{x}_0)\, \mathcal{N}(\mathbf{x}; (1-t)\mathbf{x}_0, t^2\mathbf{I})\, d\mathbf{x}_0$. In the one-dimensional Gaussian mixture studied in Figure 2, low noise levels retain separated modes and sharper changes of the score between them, whereas high noise levels produce a smoother landscape. Uniform timestep sampling exposes the student to both regimes from the start. This example motivates testing whether a gradual high-to-low noise curriculum can reduce an early preference for local modes while allowing later refinement.

(ii) Teacher sensitivity to the input timestep. The teacher s_ψ is a finite-capacity estimator and may introduce prediction errors into the DMD signal. We hypothesize that some of these errors change when the timestep supplied to the teacher is slightly increased or decreased, even with the same noisy input. If useful prediction structure is less sensitive to this perturbation, mixing the perturbed signals may reduce the influence of some errors. This hypothesis motivates our timestep offset.

(iii) Guidance composition. Writing CFG as $\tilde{s}_\psi = s_\psi(\cdot, \mathbf{c}) + w\mathbf{d}$ with $\mathbf{d} = s_\psi(\cdot, \mathbf{c}) - s_\psi(\cdot, \varnothing)$, we decompose $\mathbf{d} = \mathbf{d}_\parallel + \mathbf{d}_\perp$ relative to the conditional score. The parallel component changes the coefficient along the conditional direction, whereas the orthogonal component adds a different direction. When the parallel component is positively aligned, CFG increases the contribution along the conditional score and may produce an overly strong teacher signal. Inspired by APG (Sadat et al., 2024), we test down-weighting that component during DMD training.

These considerations motivate modifications at three locations in the same training pipeline. Our experiments evaluate their individual and combined effects on the reported video metrics.

2.3 Method Overview

Figure 1 summarizes our framework. The Self-Forcing DMD pipeline is left structurally unchanged: the student G_θ autoregressively generates chunks of latent frames under a causal KV-cache, the resulting $\hat{\mathbf{x}}_0$ is re-noised at timestep t, and a fake/teacher score pair forms the DMD gradient that

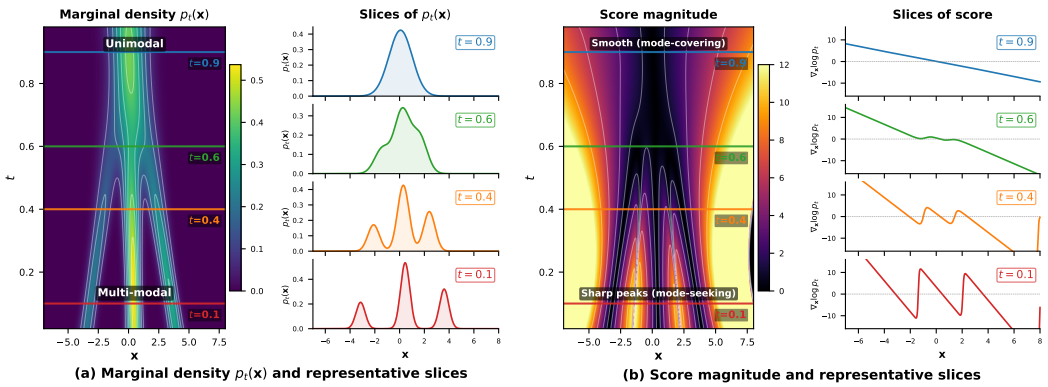


Figure 2: An illustrative 1D Gaussian mixture at different noise levels. **Left**: $p_t(\mathbf{x})$ transitions from separated modes at low t to a smoother distribution at high t. **Right**: score magnitudes illustrate variation across $\mathbf{x}$ and t; magnitude at a point is distinct from local score sharpness. This example motivates the high-to-low curriculum.

updates G_θ. Onto this pipeline we attach three lightweight, *training-only* modules that act on different parts of the training signal:

- The **timestep curriculum** (Sec. 2.4) changes *when* different noise levels are emphasized by evolving $p(t|k)$ from high-t to low-t sampling.
- The **timestep offset** (Sec. 2.5) changes *which* teacher prediction is queried by using a slightly increased or decreased input timestep $t' = t \pm \delta(t)$.
- **APG** (Sec. 2.6) changes *how* guidance is composed by reducing the weight of its parallel component while retaining the orthogonal component.

The modules can be enabled separately or combined without altering the inference procedure. Section 3 compares all single, pairwise, and full combinations under the same ablation protocol. The full combination achieves the best values on most reported metrics, supporting their joint use in the evaluated setting.

2.4 Scheduling Noise Levels via Timestep Curriculum

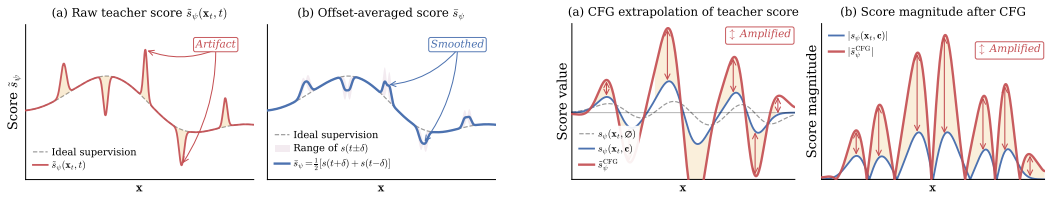
Motivation from the one-dimensional mixture. Figure 2 illustrates how noise level changes the score geometry in the mixture considered here. At high t, the mixture components overlap and the score varies relatively smoothly. At low t, the modes are separated and transitions between their regions of influence can be sharper. In standard Self-Forcing DMD, uniform timestep sampling exposes the student to low-noise signals from the beginning. We hypothesize that emphasizing higher noise levels early may reduce overly strong local attraction before gradually introducing more low-noise supervision.

Timestep curriculum. We reshape the sampling distribution $p(t|k)$ to control the order and relative frequency with which the student encounters noise levels. This changes the training distribution over t, rather than directly modifying the frozen teacher's score field. We draw t from a Beta distribution whose parameters evolve along a cosine schedule:

$$u \sim \mathrm{Beta}(\alpha_k, \beta_k), \quad t = t_{\min} + u \cdot (t_{\max} - t_{\min}), \tag{4}$$

$$\alpha_k = \alpha_0 + \rho(k)(\alpha_K - \alpha_0), \quad \beta_k = \beta_0 + \rho(k)(\beta_K - \beta_0), \quad \rho(k) = \tfrac{1}{2}\big(1 - \cos(\pi \cdot \min(k/K_w, 1))\big), \tag{5}$$

where K_w is the warmup duration. The initial parameters $(\alpha_0 > 1, \beta_0 = 1)$ bias mass toward high t, and the final parameters $(\alpha_K = 1, \beta_K > 1)$ shift it toward low t. This implements a coarse-to-fine training schedule intended to support early learning of broader structure and later refinement of detail. All interior timestep intervals retain positive sampling probability throughout training. We compare this evolving schedule with uniform sampling and fixed high- or low-noise biases in Sec. 3.4.



(a) Timestep offset smoothing.

(b) Guidance amplification illustration.

Figure 3: Conceptual illustrations motivating the teacher-side modifications. **(a)** Time perturbation may reduce the influence of prediction errors that vary with the supplied timestep. **(b)** An example of CFG amplification motivates reducing the parallel guidance weight. Our APG-based modification retains a fraction $\lambda = 0.3$ of that component. These schematics are not measurements on the video model.

2.5 Perturbing the Teacher Timestep

Motivation for timestep perturbation. Teacher prediction errors may affect the DMD gradient and contribute to visual degradation. We hypothesize that some errors change when the timestep supplied to the teacher is slightly increased or decreased, whereas useful prediction structure may remain more consistent. This motivates querying the same noisy input with perturbed timesteps, with the aim of reducing the influence of such errors on training. Figure 3 (a) illustrates this intuition.

Timestep offset. Motivated by this hypothesis, we introduce a simple augmentation: the teacher score is evaluated at a perturbed timestep $t' = t + \delta(t)$ while the fake score remains at t, with magnitude scaling as

$$\delta(t) = \delta_{\max} \cdot (t/t_{\max})^{\gamma}, \tag{6}$$

where $\gamma \geq 1$ controls the scaling profile. We apply *antithetic sampling* within each mini-batch of size B:

$$t'_i = \begin{cases} t_i + \delta(t_i), & i \leq B/2, \\ t_i - \delta(t_i), & i > B/2, \end{cases} \tag{7}$$

so that in expectation the effective teacher score becomes a local average:

$$\bar{s}_\psi(\mathbf{x}_t, t) \approx \tfrac{1}{2}\big[\tilde{s}_\psi(\mathbf{x}_t, t+\delta) + \tilde{s}_\psi(\mathbf{x}_t, t-\delta)\big], \tag{8}$$

which motivates a possible averaging effect along the timestep axis. Under the hypothesis above, this may reduce the influence of some teacher errors without substantially changing useful prediction structure. The t^γ scaling makes the perturbation larger at high noise levels and smaller at low noise levels. Combined with the curriculum, it emphasizes larger perturbations early and smaller perturbations later. We evaluate the magnitude, scaling profile, and positive/negative design empirically in Sec. 3.4.

2.6 Adjusting Guidance via Adaptive Projected Guidance

Motivation for reducing parallel guidance. As discussed in Sec. 2.2, CFG can be written as $\tilde{s}_\psi = s_\psi(\cdot, \mathbf{c}) + w(\mathbf{d}_\parallel + \mathbf{d}_\perp)$. A positively aligned parallel component increases the contribution along the conditional score and may lead to an overly strong teacher signal. Reducing its weight offers a way to adjust guidance composition while retaining the orthogonal component. This motivates our use of APG's projection mechanism during DMD training; whether it improves motion and quality is evaluated through the guidance ablation.

Adaptive Projected Guidance. We adopt the projection mechanism of APG (Sadat et al., 2024) with a weight λ on the parallel component. In the score notation used throughout this paper, the guided teacher is

$$\begin{aligned} \tilde{s}_\psi^{\mathrm{APG}} &= s_\psi(\cdot, \mathbf{c}) + w(\lambda \mathbf{d}_\parallel + \mathbf{d}_\perp), \\ \mathbf{d}_\parallel &= \frac{\langle \mathbf{d}, s_\psi(\cdot, \mathbf{c}) \rangle}{\|s_\psi(\cdot, \mathbf{c})\|^2} s_\psi(\cdot, \mathbf{c}), \qquad \mathbf{d}_\perp = \mathbf{d} - \mathbf{d}_\parallel, \end{aligned} \tag{9}$$

Table 1: Quantitative comparisons with existing methods. Our method consistently improves upon the Self-Forcing baseline across all evaluation metrics while maintaining real-time inference speed. Best results are in **bold** and second best are underlined.

Model	#Params	Resolution	Throughput↑	Latency (s)↓	Evaluation↑		
					Total	Quality	Semantic
Bidirectional Video Diffusion Models							
LTX-Video (HaCohen et al., 2024)	1.9B	768 × 512	8.98	13.5	79.83	81.88	71.62
Wan2.1 (Wan et al., 2025)	1.3B	832 × 480	0.78	103	83.37	84.30	79.65
Autoregressive Video Diffusion Models							
NOVA (Deng et al., 2025)	0.6B	768 × 480	0.88	4.1	80.31	80.66	78.92
Pyramid Flow (Jin et al., 2025)	2B	640 × 384	6.70	2.5	80.75	83.41	70.11
SkyReels-V2 (Chen et al., 2025)	1.3B	960 × 540	0.49	112	81.97	83.96	74.01
MAGI-1 (Teng et al., 2025)	4.5B	832 × 480	0.19	282	78.88	81.67	67.72
Distilled Autoregressive Video Models							
CausVid (Yin et al., 2025)	1.3B	832 × 480	17.0	0.69	81.33	83.98	70.72
Self Forcing (Huang et al., 2026)	1.3B	832 × 480	17.0	0.69	83.25	83.90	80.64
Causal Forcing (Zhu et al., 2026)	1.3B	832 × 480	17.0	0.69	84.04	84.59	**81.84**
Ours	1.3B	832 × 480	17.0	0.69	**84.16**	**85.01**	80.78



Figure 4: Frames sampled from videos generated by Self Forcing, Causal Forcing, and our method under identical prompts. Zoom in for better view.

where $\mathbf{d} = s_\psi(\cdot, \mathbf{c}) - s_\psi(\cdot, \varnothing)$. The choice $\lambda = 1$ recovers CFG, while $\lambda = 0$ retains only the orthogonal guidance component. Our default $\lambda = 0.3$ retains part of the parallel component to balance the reported motion and quality metrics. We use this projected teacher signal in the DMD update. This is intended to moderate potentially excessive parallel guidance during training.

3 Experiments

3.1 Setup

Implementation details. Following Self Forcing (Huang et al., 2026) and Causal Forcing (Zhu et al., 2026), we adopt Wan2.1-T2V-1.3B (Wan et al., 2025) as the base model, which generates 81-frame videos at a resolution of 832×480. We initialize from the publicly released Self-Forcing ODE initialization checkpoint, which corresponds to the first stage of the Self-Forcing training pipeline prior to DMD distillation, and finetune it with our three proposed modifications: timestep curriculum, teacher timestep offset, and Adaptive Projected Guidance (APG). All methods are implemented in a chunk-wise manner, where each chunk contains 3 latent frames. For the timestep curriculum (Section 2.4), we set the initial Beta parameters to $(\alpha_0, \beta_0) = (3.0, 1.0)$ and final parameters to $(\alpha_K, \beta_K) = (1.0, 3.0)$ with a cosine warmup duration of $K_w = 400$ steps. For the teacher timestep offset (Section 2.5), we use $\delta_{\max} = 0.05$ and $\gamma = 1.5$. For APG (Section 2.6), we set the parallel-component weight $\lambda = 0.3$. Training is performed for 600 steps on text prompts from VidProM (Wang & Yang, 2024) using the same protocol as Self Forcing. All experiments use 8 NVIDIA H100 GPUs.

Evaluation. We use VBench (Huang et al., 2024) as our evaluation framework and report *Total*, *Quality*, and *Semantic* scores. The main comparison (Table 1) uses the standard VBench prompt suite, while ablation studies (Sections 3.3 and 3.4) use MovieGen (Polyak et al., 2024) prompts, which are more diverse and better differentiate fine-grained design choices. For ablations, we additionally report *Dynamic Degree*, *Image Quality*, *Temporal Flickering*, *Subject Consistency*, and *Aesthetic Quality*. Training prompts (VidProM) are disjoint from both evaluation suites. We also report throughput (FPS) and latency on a single H100 GPU.

Table 2: Ablation study of proposed components. Starting from the Self-Forcing DMD baseline, we progressively add the timestep curriculum (Curri.), timestep offset (Offset), and Adaptive Projected Guidance (APG). Each component provides independent improvements, and their combination yields the best overall performance.

Curri.	Offset	APG	Quality↑	Dynamic↑	Img. Qual.↑	Flicker↑	Subject↑	Aesthetic↑
			81.22	23.12	66.35	97.93	95.60	63.19
✓			82.43	36.99	**70.08**	98.26	95.95	63.31
	✓		81.78	26.63	68.89	98.13	95.81	63.05
		✓	82.24	36.78	68.96	98.24	95.97	63.45
✓	✓		<u>82.61</u>	41.01	69.27	98.32	96.08	63.56
✓		✓	82.55	<u>45.63</u>	68.38	<u>98.42</u>	<u>96.18</u>	<u>63.64</u>
	✓	✓	82.48	41.31	69.19	98.35	96.12	63.49
✓	✓	✓	**82.90**	**56.48**	**70.08**	**98.65**	**96.22**	**63.72**



Figure 5: **Qualitative ablation.** Each column shows frames generated under the same prompt by progressively removing one component from the full model: *w/o APG*, *w/o Offset*, *w/o Curriculum*, and the *Full Model*. Best viewed zoomed in.

3.2 Results

Performance comparison with existing models. We compare our method against baselines of comparable scale, including bidirectional video diffusion models Wan2.1 (Wan et al., 2025) and LTX-Video (HaCohen et al., 2024), autoregressive video diffusion models NOVA (Deng et al., 2025), Pyramid Flow (Jin et al., 2025), SkyReels-V2 (Chen et al., 2025), and MAGI-1 (Teng et al., 2025), as well as distilled autoregressive video models CausVid (Yin et al., 2025), Self Forcing (Huang et al., 2026), and Causal Forcing (Zhu et al., 2026). As shown in Table 1, our method obtains the highest reported Total and Quality scores among the compared methods, with throughput of 17.0 FPS and latency of 0.69s. Relative to Self Forcing, Total, Quality, and Semantic improve by 0.91, 1.11, and 0.14 points, respectively. Relative to Causal Forcing, Total and Quality are higher by 0.12 and 0.42 points, whereas Semantic is lower by 1.06 points. This comparison indicates a quality–semantic trade-off in the reported results, while showing the benefits of modifying the training signal at the same inference speed.

Qualitative comparison. Figure 4 presents frames generated by our method and the key baselines under identical prompts. In these examples, Self Forcing exhibits softer textures and less distinct fine details, while Causal Forcing improves some structures but retains artifacts around certain objects. Our method produces sharper details and better-preserved textures in the displayed comparisons. These examples illustrate the visual differences alongside the quantitative metrics. Please refer to the supplementary material for full video results.

3.3 Ablation Study

To evaluate the contribution of each component, we conduct an ablation study starting from the Self-Forcing DMD baseline and adding each intervention individually and in combination. All ablation experiments use the same training budget and evaluation protocol. As shown in Table 2, each component improves Quality and Dynamic Degree over the baseline. The full combination achieves the highest Quality, Dynamic Degree, Temporal Flickering, Subject Consistency, and Aesthetic scores, and ties the curriculum-only configuration on Image Quality. Its Dynamic Degree is 56.48, compared with 23.12 for the baseline and 45.63 for the best pairwise configuration; Quality is 82.90, compared with 82.61 for the best pairwise configuration. These results support combining the interventions in

Table 3: Effect of timestep curriculum design. "High→Low" denotes our proposed curriculum that transitions from high-t to low-t bias. K_w is the warmup duration in training steps. The high-to-low curriculum with $K_w = 400$ achieves the best balance between Dynamic Degree and visual quality.

Schedule	Quality↑	Dynamic↑	Img. Qual.↑	Flicker↑	Subject↑	Aesthetic↑
Uniform (α=1, β=1)	81.22	23.12	66.35	97.93	95.60	63.19
High-t only (α=3, β=1, fixed)	81.85	31.42	67.53	97.78	95.43	62.91
Low-t only (α=1, β=3, fixed)	80.94	19.85	65.18	98.05	95.32	62.74
High→Low, $K_w = 200$	82.18	33.56	69.21	98.12	95.78	63.22
High→Low, $K_w = 400$ (ours)	**82.43**	**36.99**	**70.08**	**98.26**	**95.95**	63.31
High→Low, $K_w = 600$	82.35	35.74	69.85	98.21	95.89	**63.38**

Table 4: Effect of timestep offset design. $\delta_{\max}$ controls the maximum perturbation magnitude, γ controls the scaling profile, and "Antithetic" indicates whether antithetic sampling is used. Our default setting ($\delta_{\max}$=0.05, γ=1.5, antithetic) achieves the best performance.

$\delta_{\max}$	γ	Antithetic	Quality↑	Dynamic↑	Img. Qual.↑	Flicker↑	Subject↑	Aesthetic↑
0 (none)	–	–	81.22	23.12	66.35	97.93	95.60	**63.19**
0.02	1.5	✓	81.48	24.86	67.52	98.01	95.68	63.11
0.05	1.5	✓	**81.78**	**26.63**	**68.89**	**98.13**	**95.81**	63.05
0.10	1.5	✓	81.35	25.18	67.14	97.96	95.55	62.88
0.05	1.0	✓	81.62	25.91	68.23	98.09	95.77	62.97
0.05	2.0	✓	81.68	26.14	68.51	98.07	95.74	63.08
0.05	1.5		81.51	24.53	67.78	97.98	95.65	62.93

the evaluated setting. Dynamic Degree quantifies motion intensity; we interpret its increase alongside the quality and temporal consistency metrics.

Qualitative Ablation. Figure 5 shows frames generated under the same prompt after removing each component from the full model. In the displayed examples, *without APG*, color saturation is less consistent across parts of the scene (yellow boxes); *without Offset*, some textures and fine details appear softer; and *without Curriculum*, subjects show less visible change across the sampled frames (red boxes). The full model combines more consistent colors, clearer textures, and visible motion in these examples. These comparisons illustrate the observed effects of the components alongside the quantitative ablation.

3.4 Analysis of Design Choices

Beyond evaluating component contributions, we analyze the key design choices within the current training and evaluation setting.

Timestep curriculum design. Table 3 compares uniform sampling, fixed high-t bias, fixed low-t bias, and high-to-low curricula with different warmup durations K_w. Fixed high-t sampling increases Dynamic Degree relative to uniform sampling, but its Flicker and Subject Consistency scores are lower. Fixed low-t sampling yields lower Quality and Dynamic Degree than uniform sampling. All three tested high-to-low schedules improve those two metrics, with $K_w = 400$ achieving the highest Quality and Dynamic Degree among the tested durations. The $K_w = 600$ configuration has a slightly higher Aesthetic score. These observations support a changing sampling distribution for this budget and are consistent with the curriculum motivation, without establishing a universally optimal warmup duration.

Timestep offset design. Table 4 varies the offset magnitude $\delta_{\max}$, scaling exponent γ, and antithetic sampling. Among the tested magnitudes, $\delta_{\max} = 0.05$ gives the highest Quality and Dynamic Degree; both smaller and larger offsets yield smaller gains. With that magnitude, $\gamma = 1.5$ gives the highest Quality and Dynamic Degree among the tested profiles. The antithetic configuration also scores higher than the one-sided offset on the reported metrics. These comparisons support the selected perturbation design in this setting. Aesthetic Quality remains slightly below the no-offset baseline for all tested offset configurations.

Table 5: Comparison of guidance techniques for the teacher signal in DMD training. The parallel-component weight $\lambda = 0.3$ gives the highest reported quality metrics among these settings, while smaller weights yield higher Dynamic Degree.

Guidance	Quality↑	Dynamic↑	Img. Qual.↑	Flicker↑	Subject↑	Aesthetic↑
No guidance (w=0)	80.15	**42.37**	63.28	97.52	95.11	62.18
CFG++ (Saini et al., 2026)	81.08	25.46	65.87	97.85	95.48	63.02
APG, $\lambda = 0.0$	81.45	<u>39.82</u>	67.14	97.68	95.35	63.28
APG, $\lambda = 0.3$ (ours)	**82.24**	36.78	**68.96**	**98.24**	**95.97**	**63.45**
APG, $\lambda = 0.7$	<u>81.72</u>	28.95	<u>67.85</u>	<u>98.08</u>	<u>95.78</u>	<u>63.35</u>
APG, $\lambda = 1.0$ (=CFG)	81.22	23.12	66.35	97.93	95.60	63.19

Guidance technique comparison. Table 5 compares no guidance, standard CFG, the rectified-flow CFG++ variant (Saini et al., 2026), and APG with varying parallel-component weight λ. Without guidance, Dynamic Degree is highest, while Quality and the other reported quality metrics are lower. Pure orthogonal guidance ($\lambda = 0$) increases Dynamic Degree relative to CFG, but does not give the best quality metrics. The default $\lambda = 0.3$ attains the highest Quality, Image Quality, Flicker, Subject Consistency, and Aesthetic scores among the compared guidance settings. Dynamic Degree decreases as λ increases from 0 to 1, while the other APG metrics peak at 0.3 among the tested values. The results favor partial parallel-component retention as a motion–quality compromise in this setting, rather than complete removal as a universal choice.

4 Related Work

Bidirectional video diffusion models. Most state-of-the-art video generators adopt a bidirectional denoising paradigm, where all frames are jointly denoised over multiple diffusion steps (Wan et al., 2025; HaCohen et al., 2024). These models achieve high visual quality and temporal coherence but require loading the entire video into memory, resulting in latency and memory costs that scale with video length, making them impractical for interactive or streaming applications.

Autoregressive video generation and distillation. Autoregressive approaches generate video frames or chunks causally, conditioning each on previously generated outputs Guo et al. (2026); Wu et al. (2026); Yi et al. (2025); Yesiltepe et al. (2025); Cui et al. (2025); Wu et al. (2025); Li et al. (2025); Ma et al. (2026). Self-Forcing (Huang et al., 2026) closes the train-test gap by training the student on its own outputs and combines this with Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) for real-time generation. Causal Forcing (Zhu et al., 2026) further improves quality by refining the ODE initialization strategy. Our work addresses the DMD training signal through timestep sampling, teacher timestep perturbation, and guidance composition, motivated by a score sharpness perspective. This offers a training-signal perspective alongside approaches that improve ODE initialization.

5 Conclusion

We revisited Self-Forcing DMD from a *score sharpness* perspective to address insufficient motion and degraded visual quality in autoregressive video generation. An illustrative one-dimensional analysis motivates a timestep curriculum, while hypotheses about teacher prediction sensitivity and guidance composition motivate a timestep offset and an APG-based modification. These interventions act on the training signal without changing the student architecture or inference procedure. In the reported experiments, our method improves Total, Quality, and Semantic scores over Self-Forcing and obtains the highest Total and Quality scores among the compared methods. Relative to Causal Forcing, the gains in Quality and Total accompany a lower Semantic score. Component ablations support combining the interventions, with the full model attaining the best values on most reported metrics while retaining the same real-time inference speed.

Limitations and future work. This work focuses on training signal design within Self-Forcing DMD. Extending these interventions to other distillation frameworks is a promising direction. We also use a predefined timestep curriculum; future work could adapt timestep sampling to the training state to reduce manual configuration and accommodate different training budgets.

AI USE STATEMENT

In this work, we used generative AI tools to edit parts of the paper to improve readability. We have reviewed all AI-assisted parts. Specifically, we checked LLM-generated text edits for accuracy and consistency with our original content. We take full responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work improves the quality of real-time autoregressive video generation, which has broad potential applications in creative content production, education, interactive entertainment, and assistive design tools. By enabling higher-fidelity video synthesis at real-time speeds, our method can make video generation more accessible to a wider audience of creators and developers.

Our work does not involve the collection of human data and all training data (VidProM text prompts and the pre-trained Wan2.1 model) are publicly available. The proposed modifications are training-only and do not introduce new capabilities beyond improving the quality of an existing generation paradigm.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. Detailed implementation information is provided in the paper: the overall training algorithm and the integration of our three proposed modules (timestep curriculum, timestep offset, and Adaptive Projected Guidance) are described in the Method Section (Section 2) and formalized in Algorithm 1 in the Appendix. Complete hyperparameters, model architecture, optimization settings and hardware configuration are listed in Table 6 in the Appendix. The evaluation protocol, metrics, and dataset are specified in Section 3 and Appendix C. All baseline results are either taken from official publications or reproduced using publicly released checkpoints with default inference settings. Formal analyses supporting our claims, including score sharpness in Gaussian mixtures, bias cancellation of antithetic offset and magnitude bounds for APG, are provided in Appendix E.

REFERENCES

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, volume 2025, pp. 44730–44745, 2025.

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.

Hang Guo, Zhaoyang Jia, Jiahao Li, Bin Li, Yuanhao Cai, Jiangshan Wang, Yawei Li, and Yan Lu. Efficient autoregressive video diffusion with dummy head. *arXiv preprint arXiv:2601.20499*, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Jing Jia, Sifan Liu, Bowen Song, Wei Yuan, Liyue Shen, and Guanyang Wang. Antithetic noise in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/20b6b87ca17792337f414d948af7b0e8-Abstract-Conference.html.

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. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025.

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

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.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Yuexiao Ma, Xuzhe Zheng, Jing Xu, Xiwei Xu, Feng Ling, Xiawu Zheng, Huafeng Kuang, Huixia Li, Xing Wang, Xuefeng Xiao, et al. Flow caching for autoregressive video generation. *arXiv preprint arXiv:2602.10825*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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.

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*, pp. 10684–10695, 2022.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Shreshth Saini, Shashank Gupta, and Alan Bovik. Rectified cfg++ for flow based models. *Advances in Neural Information Processing Systems*, 38:149034–149074, 2026.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026.

Xiaofei Wu, Guozhen Zhang, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Xuming He. Pack and force your memory: Long-form and consistent video generation. *arXiv preprint arXiv:2510.01784*, 2025.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Extended Related Work

A.1 Bidirectional Video Diffusion Models

Diffusion-based (Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2022; Ma et al., 2024; Liu et al., 2022; Esser et al., 2024; Peebles & Xie, 2023) video generation has advanced rapidly in recent years. Most state-of-the-art systems adopt a bidirectional denoising paradigm, where all frames are jointly denoised over multiple diffusion steps. Wan2.1 (Wan et al., 2025) scales transformer-based architectures for high-fidelity text-to-video synthesis. LTX-Video (HaCohen et al., 2024) introduces latent diffusion in the temporal domain for efficient generation. Movie Gen (Polyak et al., 2024) demonstrates large-scale video generation with foundation model capabilities. While these models achieve impressive visual quality and temporal coherence, they require loading the entire video into memory during denoising, resulting in latency and memory costs that scale with video length and making them impractical for interactive or streaming applications.

A.2 Autoregressive Video Generation and Distillation

Autoregressive approaches generate video frames or chunks causally, conditioning each on previously generated outputs. NOVA (Deng et al., 2025) pioneers continuous-valued autoregressive video generation without vector quantization. Pyramid Flow (Jin et al., 2025) introduces a pyramidal architecture for efficient multi-scale generation. SkyReels-V2 (Chen et al., 2025) and MAGI-1 (Teng et al., 2025) extend autoregressive generation to longer sequences through hierarchical chunk-based architectures. Several concurrent works further explore this direction: Geometry Forcing (Wu et al., 2026) enforces 3D consistency in autoregressive video diffusion, Deep Forcing (Yi et al., 2025) addresses long-range coherence through deep sink mechanisms, and Infinity-ROPE (Yesiltepe et al., 2025) enables action-controllable infinite generation via self-rollout.

Self-Forcing (Huang et al., 2026) closes the train-test gap by training the student on its own outputs and combines this with Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) for real-time generation. CausVid (Yin et al., 2025) demonstrates that bidirectional video diffusion models can be converted to fast autoregressive generators through causal distillation. Causal Forcing (Zhu et al., 2026) further improves quality by refining the ODE initialization strategy with a noise-refreshing mechanism. Self-Forcing++ (Cui et al., 2025) extends the paradigm to minute-scale generation, while Pack-and-Force (Wu et al., 2025) and Stable Video Infinity (Li et al., 2025) address long-form consistency through memory-efficient designs. Flow Caching (Ma et al., 2026) accelerates autoregressive video diffusion by reusing intermediate flow computations. Our work targets the DMD training signal through timestep sampling, teacher timestep perturbation, and guidance composition, offering a score sharpness perspective alongside these initialization- and architecture-based approaches.

A.3 Diffusion Distillation

Diffusion distillation aims to reduce the number of sampling steps required by pre-trained diffusion models. Progressive Distillation (Salimans & Ho, 2022) halves the number of steps iteratively by training a student to match two teacher steps in one. Consistency Models (Song et al., 2023) enforce self-consistency along the ODE trajectory, enabling single-step generation. Distribution Matching Distillation (DMD) (Yin et al., 2024b) takes a different approach by minimizing the KL divergence between student and data distributions through a learned fake score network, and DMD2 (Yin et al., 2024a) improves this with a regression loss and two-timescale update rule. Our work builds on DMD and studies training interventions for insufficient motion and visual degradation in autoregressive Self-Forcing generation, motivated by a score sharpness perspective.

A.4 Guidance Techniques for Diffusion Models

Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) enhances text-conditional generation by combining conditional and unconditional predictions. Its strength can affect the balance between sample quality and variation. The rectified-flow CFG++ variant (Saini et al., 2026) studies guidance for flow-based models. Adaptive Projected Guidance (APG) (Sadat et al., 2024) decomposes the guidance direction into parallel and orthogonal components and combines this with rescaling and

momentum in its original formulation. We adopt its projection mechanism as a *training-time* intervention in DMD, down-weighting the parallel component relative to the conditional score. Our default retains part of this component, as specified in Eq. 9, and the guidance ablation evaluates the resulting motion–quality trade-off.

Antithetic Noise in Diffusion Models (Jia et al., 2026) studies paired initial noise samples and their negations for diffusion sampling and uncertainty estimation. Our offset instead perturbs the teacher's input timestep during distillation while retaining the same noisy input. The two approaches use symmetry at different locations in the generation and training pipelines.

B Implementation Details

B.1 Training Configuration

We provide the complete training configuration in Table 6. All experiments are conducted on 8 NVIDIA H100 GPUs.

B.2 Training Algorithm

We summarize the complete training procedure of our method in Algorithm 1. The algorithm highlights how the three proposed modifications (timestep curriculum, timestep offset, and APG) are integrated into the standard Self-Forcing DMD training loop. All three modules are lightweight and only modify the training signal without changing the model architecture or inference pipeline.

The algorithm has a two-phase structure following the DMD protocol (Yin et al., 2024a). **Phase 1** (every step) updates the fake score network s_ϕ via denoising score matching on student-generated samples, keeping s_ϕ an accurate density estimator of the current student distribution. **Phase 2** (every R=5 steps) updates the generator G_θ using the DMD gradient—the difference between s_ϕ and the APG-enhanced teacher $\tilde{s}_\psi^{\mathrm{APG}}$. Our three modifications (curriculum, offset, APG) only affect Phase 2; the fake score training in Phase 1 remains unchanged from standard DMD. Note that the antithetic sampling for timestep offset is applied across the mini-batch (first half uses $t' = t + \delta$, second half uses $t' = t - \delta$), so that in expectation the teacher signal approximates the local average in Eq. 8.

B.3 Inference Configuration

At inference time, the student model generates videos autoregressively in a chunk-wise manner using the causal KV-cache inherited from Self-Forcing. Our three proposed modifications are *training-only* and introduce zero overhead at inference: the timestep curriculum only affects the training-time sampling distribution, the timestep offset only perturbs the teacher evaluation during training, and APG only modifies the teacher score computation in the DMD gradient. The inference pipeline is therefore identical to the standard Self-Forcing DMD pipeline, achieving a throughput of 17.0 FPS and a latency of 0.69 seconds on a single NVIDIA H100 GPU.

C Evaluation Details

C.1 Evaluation Metrics

We use VBench (Huang et al., 2024) as our primary evaluation framework. VBench provides a comprehensive suite of metrics organized into two categories:

- **Quality** measures the visual fidelity and temporal coherence of generated videos, aggregating several sub-metrics such as *Image Quality*(per-frame visual fidelity assessed by a learned quality predictor), *Temporal Flickering*(inter-frame consistency measured by pixel-level and feature-level differences between consecutive frames. Higher values indicate less flickering), *Subject Consistency*(identity preservation of the main subject across frames, measured by DINO feature similarity), *Aesthetic Quality*(per-frame aesthetic appeal assessed by an aesthetic scoring model) and *Dynamic Degree*(the amount of motion in a video, used here alongside quality and temporal consistency metrics).

Table 6: Complete training hyperparameters.

Hyperparameter	Value
Base Model	
Architecture	Wan2.1-T2V-1.3B (student generator)
Teacher (real score)	Wan2.1-T2V-14B (frozen)
Resolution	832×480
Video length	81 frames (21 latent frames)
Chunk size	3 latent frames
Optimization	
Initialization	Self-Forcing ODE init. checkpoint
Optimizer	AdamW (β_1=0.0, β_2=0.999)
Generator learning rate	2.0×10^{-6}
Fake score learning rate	4.0×10^{-7}
Weight decay	0.01
Fake/generator update ratio	5
Batch size per GPU	1
Total batch size	64
Gradient accumulation steps	8 (for 8-GPU setup)
Training steps	600
Mixed precision	Yes (bf16)
Gradient checkpointing	Yes
EMA weight	0.99
EMA start step	200
Hardware	8× NVIDIA H100 (80GB) GPUs
Training prompts	VidProM (Wang & Yang, 2024)
Diffusion Settings	
Guidance scale w	3.0
Timestep shift	5.0
Denoising steps	[1000, 750, 500, 250]
Timestep Curriculum (Sec. 2.4)	
Initial Beta params (α_0, β_0)	$(3.0, 1.0)$
Final Beta params (α_K, β_K)	$(1.0, 3.0)$
Warmup duration K_w	400 steps
Annealing schedule	Cosine
Timestep Offset (Sec. 2.5)	
Max offset $\delta_{\max}$	0.05
Scaling exponent γ	1.5
Antithetic sampling	Yes
APG (Sec. 2.6)	
Parallel-component weight λ	0.3

- **Semantic** measures the alignment between the generated video and the input text prompt, including object identity, action correctness, spatial relationships, and scene composition.

The **Total** score is a weighted combination of Quality and Semantic scores as defined by the VBench benchmark.

C.2 Evaluation Protocol

For the main comparison (Table 1), we use the standard VBench prompt suite to ensure fair comparison with published results. For ablation studies (Sections 3.3 and 3.4), we use prompts from MovieGen (Polyak et al., 2024), which are more diverse and better differentiate fine-grained design

Algorithm 1 Self-Forcing DMD Training with Score Sharpness Mitigation

Require: Student generator G_θ, fake score network s_ϕ, frozen teacher score s_ψ
Require: Curriculum params: $(\alpha_0, \beta_0, \alpha_K, \beta_K, K_w)$; Offset params: $(\delta_{\max}, \gamma)$; APG param: λ; Guidance scale: w; Update ratio: R

for training step $k = 1, 2, \ldots, K$ **do**
  // — Phase 1: Fake Score Update (every step) —
  $\hat{\mathbf{x}}_0^{(\text{critic})} \leftarrow G_\theta(\boldsymbol{\epsilon}, \mathbf{c})$ with `stop_gradient` ▷ Generate samples (no grad to G_θ)
  Sample $t_c \sim \text{Uniform}, \quad \boldsymbol{\epsilon}_c \sim \mathcal{N}(0, \mathbf{I})$
  $\mathbf{x}_{t_c} \leftarrow (1 - t_c)\,\hat{\mathbf{x}}_0^{(\text{critic})} + t_c\,\boldsymbol{\epsilon}_c$ ▷ Re-noise generated sample
  $\hat{\mathbf{x}}_0^{(\text{pred})} \leftarrow s_\phi(\mathbf{x}_{t_c}, t_c, \mathbf{c})$ ▷ Fake score denoising prediction
  $\mathcal{L}_{\text{critic}} \leftarrow \|\hat{\mathbf{x}}_0^{(\text{pred})} - \hat{\mathbf{x}}_0^{(\text{critic})}\|^2$ ▷ Denoising score matching loss
  Update s_ϕ using $\nabla_\phi \mathcal{L}_{\text{critic}}$

  if $k \bmod R = 0$ **then**
    // — Phase 2: Generator Update (every R steps) —

    // Self-Forcing: generate video autoregressively
    $\hat{\mathbf{x}}_0 \leftarrow G_\theta(\boldsymbol{\epsilon}, \mathbf{c})$ ▷ Autoregressive generation with causal KV-cache

    // Timestep Curriculum (Sec. 2.4): evolving Beta distribution
    $\rho \leftarrow \frac{1}{2}(1 - \cos(\pi \cdot \min(k/K_w, 1)))$
    $\alpha_k \leftarrow \alpha_0 + \rho(\alpha_K - \alpha_0), \quad \beta_k \leftarrow \beta_0 + \rho(\beta_K - \beta_0)$
    $u \sim \text{Beta}(\alpha_k, \beta_k), \quad t \leftarrow t_{\min} + u \cdot (t_{\max} - t_{\min})$

    // Re-noise and compute scores
    $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad \mathbf{x}_t \leftarrow (1 - t)\,\hat{\mathbf{x}}_0 + t\,\boldsymbol{\epsilon}$

    // Timestep Offset (Sec. 2.5): antithetic perturbation for teacher
    $\delta \leftarrow \delta_{\max} \cdot (t/t_{\max})^\gamma$ ▷ Scaled offset
    $t' \leftarrow t + \delta$ (first half of batch), $\quad t' \leftarrow t - \delta$ (second half) ▷ Antithetic

    // APG (Sec. 2.6): projected guidance for teacher score
    $\mathbf{s}_c \leftarrow s_\psi(\mathbf{x}_t, t', \mathbf{c}), \quad \mathbf{s}_\varnothing \leftarrow s_\psi(\mathbf{x}_t, t', \varnothing)$ ▷ Teacher at perturbed t'
    $\mathbf{d} \leftarrow \mathbf{s}_c - \mathbf{s}_\varnothing, \quad \mathbf{d}_\perp \leftarrow \mathbf{d} - \frac{\langle \mathbf{d}, \mathbf{s}_c \rangle}{\|\mathbf{s}_c\|^2}\mathbf{s}_c, \quad \mathbf{d}_\parallel \leftarrow \mathbf{d} - \mathbf{d}_\perp$
    $\tilde{s}_\psi^{\text{APG}} \leftarrow \mathbf{s}_c + w \cdot (\lambda \cdot \mathbf{d}_\parallel + \mathbf{d}_\perp)$

    // DMD gradient
    $\hat{s}_\phi \leftarrow s_\phi(\mathbf{x}_t, t, \mathbf{c})$ ▷ Fake score at original t (no offset)
    $\Delta \leftarrow \hat{s}_\phi - \tilde{s}_\psi^{\text{APG}}$
    Update G_θ using $\nabla_\theta \mathcal{L}_{\text{gen}} \propto \Delta \cdot \frac{\partial \hat{\mathbf{x}}_0}{\partial \theta}$
    Update EMA of G_θ (if $k \geq k_{\text{ema_start}}$)
  end if
end for
return Trained student generator G_θ (EMA weights)

choices. Training prompts are drawn from VidProM (Wang & Yang, 2024) and are strictly disjoint from both evaluation suites, ensuring no train-test contamination.

All baseline results are either taken from official publications or reproduced using publicly released checkpoints with default inference settings. Throughput (FPS) and latency are measured on a single NVIDIA H100 GPU.

D Additional Qualitative Results

We provide additional qualitative comparisons in Figure 6 and 7. For each prompt, we show representative frames generated by Self Forcing, Causal Forcing, and our method under identical conditions. These samples are selected without cherry-picking to reflect the typical output quality of each method. Consistent with the quantitative results in the main paper, our method generates videos with sharper textures, more faithful colors, and richer motion compared to the baselines. We encourage readers to view the full video results on our anonymous project page for a more comprehensive assessment of temporal coherence and motion quality. **A video containing qualitative comparisons is also attached to the supplementary materials.**

E Formal Analysis of Score Sharpness

In this section, we analyze the one-dimensional Gaussian mixture used to motivate the curriculum, discuss the symmetric timestep average underlying the offset intuition, and derive a conditional magnitude relation for projected guidance. These calculations clarify the design perspective introduced in Section 2.2.

E.1 Score Sharpness in Gaussian Mixtures

Consider a 1D Gaussian mixture with M equally weighted components:

$$p_{\text{data}}(x) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{N}(x;\mu_m,\sigma_0^2). \tag{10}$$

Under the rectified flow interpolation $x_t = (1-t)x_0 + t\epsilon$ with $\epsilon \sim \mathcal{N}(0,1)$, the marginal at noise level t is:

$$p_t(x) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{N}\big(x;\,(1-t)\mu_m,\,(1-t)^2\sigma_0^2 + t^2\big). \tag{11}$$

Each component has mean $(1-t)\mu_m$ and variance $\sigma_t^2 = (1-t)^2\sigma_0^2 + t^2$. The score function is:

$$\nabla_x \log p_t(x) = \sum_{m=1}^{M} w_m(x,t)\cdot\frac{(1-t)\mu_m - x}{\sigma_t^2}, \tag{12}$$

where $w_m(x,t) = \frac{\mathcal{N}(x;(1-t)\mu_m,\sigma_t^2)}{\sum_{m'}\mathcal{N}(x;(1-t)\mu_{m'},\sigma_t^2)}$ are the posterior responsibilities.

Noise level and local score variation. Local score sharpness can be characterized by $|\nabla_x^2 \log p_t(x)|$. A uniform bound on this quantity over a domain yields a Lipschitz bound on the score over that domain. In the single-component regime (when x is close to one mode), the score reduces to $\nabla_x \log p_t(x) \approx \frac{(1-t)\mu_m - x}{\sigma_t^2}$, whose slope has magnitude $1/\sigma_t^2$. We have:

$$\sigma_t^2 = (1-t)^2\sigma_0^2 + t^2 \xrightarrow{t\to 0} \sigma_0^2, \qquad \sigma_t^2 \xrightarrow{t\to 1} 1. \tag{13}$$

When the data variance $\sigma_0^2 \ll 1$ (as is typical for well-clustered modes), the score curvature $1/\sigma_t^2$ is much larger at small t than at large t. More importantly, the inter-component separation $(1-t)(\mu_m - \mu_{m'})$ shrinks as $t \to 1$, causing the posterior weights w_m to become nearly uniform (all components contribute equally), yielding a smooth, unimodal score. As $t \to 0$, the components separate and $w_m(x,t)$ concentrates on the nearest mode, producing sharp transitions in the score landscape between modes. This contrast in the illustrative mixture motivates the high-to-low curriculum: early training emphasizes higher noise levels, followed by increased low-noise supervision for detail refinement.

E.2 Bias Cancellation of Antithetic Offset

Proposition. *Let $s_\psi(\mathbf{x}_t, t)$ be the teacher score evaluated at noise level t. For any differentiable s_ψ and offset $\delta > 0$, the antithetic average satisfies:*

$$\bar{s}_\psi(\mathbf{x}_t,t) \triangleq \frac{1}{2}\big[s_\psi(\mathbf{x}_t,t+\delta) + s_\psi(\mathbf{x}_t,t-\delta)\big] = s_\psi(\mathbf{x}_t,t) + \frac{\delta^2}{2}\frac{\partial^2 s_\psi}{\partial t^2}\bigg|_t + O(\delta^4). \tag{14}$$

Proof. By Taylor expansion around t:

$$s_\psi(\mathbf{x}_t, t+\delta) = s_\psi(\mathbf{x}_t, t) + \delta\frac{\partial s_\psi}{\partial t} + \frac{\delta^2}{2}\frac{\partial^2 s_\psi}{\partial t^2} + O(\delta^3), \tag{15}$$

$$s_\psi(\mathbf{x}_t, t-\delta) = s_\psi(\mathbf{x}_t, t) - \delta\frac{\partial s_\psi}{\partial t} + \frac{\delta^2}{2}\frac{\partial^2 s_\psi}{\partial t^2} + O(\delta^3). \tag{16}$$

Averaging eliminates the first-order term $\delta\frac{\partial s_\psi}{\partial t}$, leaving only the second-order correction. □

Interpretation. The symmetric average motivates a possible averaging effect when the teacher is queried with slightly larger and smaller timesteps. We hypothesize that some prediction errors vary more with these changes than useful prediction structure, so such averaging may reduce their influence on training. This is the mechanism intuition behind the offset; the expansion alone does not identify which parts of a prediction are errors.

A *unidirectional* offset queries only one side of the original timestep. The symmetric design explores both directions and is compared with a one-sided perturbation in Table 4. The reported results favor the symmetric choice in the evaluated setting.

E.3 Geometry of Projected Guidance

Proposition. *Let $\mathbf{s}_c \neq 0$ be the conditional score, $\mathbf{d}$ the guidance direction, and $w > 0$. Define the signed projection coefficient $\alpha = \langle \mathbf{d}, \mathbf{s}_c\rangle / \|\mathbf{s}_c\|^2$, so that $\mathbf{d}_\parallel = \alpha\mathbf{s}_c$ and $\mathbf{d}_\perp = \mathbf{d} - \alpha\mathbf{s}_c$. For $0 \leq \lambda \leq 1$, let*

$$\mathbf{g}_\lambda = \mathbf{s}_c + w(\lambda\mathbf{d}_\parallel + \mathbf{d}_\perp). \tag{17}$$

Then its squared magnitude is

$$\|\mathbf{g}_\lambda\|^2 = (1 + w\lambda\alpha)^2\|\mathbf{s}_c\|^2 + w^2\|\mathbf{d}_\perp\|^2. \tag{18}$$

If $\alpha \geq 0$, decreasing λ cannot increase this magnitude. In particular, $\|\mathbf{g}_\lambda\| \leq \|\mathbf{g}_1\|$, where $\mathbf{g}_1$ is CFG. For $\lambda < 1$, equality under these assumptions holds exactly when $\alpha = 0$.

Proof. Substituting $\mathbf{d}_\parallel = \alpha\mathbf{s}_c$ gives $\mathbf{g}_\lambda = (1 + w\lambda\alpha)\mathbf{s}_c + w\mathbf{d}_\perp$. Since $\mathbf{s}_c \perp \mathbf{d}_\perp$, the squared norm separates into the two terms above. When $\alpha \geq 0$, both $1 + w\lambda\alpha$ and its square are nondecreasing in $\lambda \in [0, 1]$, while the orthogonal term is unchanged. This gives the stated inequality and equality condition. □

Interpretation for DMD. The identity distinguishes the parallel coefficient from the orthogonal contribution and applies to the default $\lambda = 0.3$ as well as the pure orthogonal case $\lambda = 0$. For positively aligned parallel guidance, down-weighting this component moderates the teacher signal's magnitude while retaining $w\mathbf{d}_\perp$. This provides a conditional geometric motivation for the intervention. With negatively aligned parallel guidance, the magnitude need not decrease. The relation concerns the teacher signal; the DMD discrepancy also depends on the fake score. Its effect on video quality and motion is evaluated through the guidance ablation.

F Broader Impact

Our work improves the quality of real-time autoregressive video generation, which has broad potential applications in creative content production, education, interactive entertainment, and assistive design tools. By enabling higher-fidelity video synthesis at real-time speeds, our method can make video generation more accessible to a wider audience of creators and developers.

Our work does not involve the collection of human data, and all training data (VidProM text prompts and the pre-trained Wan2.1 model) are publicly available. The proposed modifications are training-only and do not introduce new capabilities beyond improving the quality of an existing generation paradigm.



Figure 6: More qualitative results of our method. Zoom in for better view.



Figure 7: More qualitative results of our method. Zoom in for better view.