ON-POLICY ADVERSARIAL FLOW DISTILLATION FOR AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video generators are attractive for streaming, long-horizon, and interactive applications, but distilling strong black-box teachers into causal students remains difficult. The student must learn under its own rollout distribution, whereas practical teachers may expose only prompt-conditioned completed videos and may differ in architecture, capacity, temporal design, and sampling schedule. This interface makes supervised fine-tuning off-policy, score-based distillation inapplicable, and direct adversarial imitation too sparse for denoising-time credit assignment. We propose *Adversarial Flow Distillation* (AFD), an on-policy framework for heterogeneous black-box video distillation. AFD trains a prompt-paired Bradley–Terry discriminator on teacher videos and current student rollouts, then uses its rollout-specific advantage to weight forward-process flow-matching updates on the student's noised states. This provides dense velocity supervision without teacher scores, denoising trajectories, or step alignment. At the population optimum, the discriminator score recovers the teacher through exponential tilting, while its centered linear tilt yields strict local reverse-KL descent. Across two causal AR student families, AFD improves motion- and physics-sensitive generation while preserving general video quality. Independent physics evaluation and feedback ablations further support these gains and the importance of adaptive on-policy feedback. The method requires only clean teacher videos and student rollouts, providing a practical route for distilling proprietary or heterogeneous video generators into efficient autoregressive students.

1 INTRODUCTION

Efficient deployment of modern video generators increasingly depends on distilling large diffusion, score-based, or flow-matching models into smaller students (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Liu et al., 2022; Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024; Kim et al., 2025; Chern et al., 2025). In practice, many strong video teachers are accessible only as black-box samplers that return completed clips, without exposing scores, logits, latents, sampler states, or denoising trajectories (Brooks et al., 2024; Polyak et al., 2024; Kling Team, 2025). This is especially restrictive for autoregressive (AR) students such as Self-Forcing (Huang et al., 2025): a black-box teacher may use a different architecture, temporal conditioning scheme, and long denoising schedule, while the deployable AR student generates causally with few denoising steps. Existing distillation recipes therefore lack an aligned supervision channel from the teacher's hidden generation process to the student's rollout distribution and intermediate noised states.

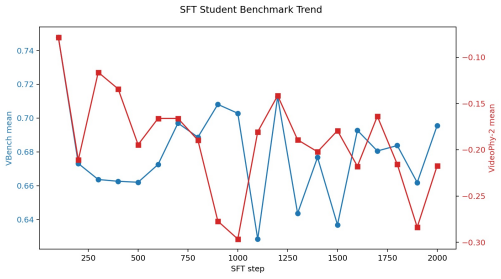

Figure 1: Preliminary SFT trend for an AR video student. VBench and VideoPhy-2 scores decrease rather than improve monotonically as SFT steps increase.

A straightforward adaptation strategy is supervised fine-tuning (SFT) on teacher-generated videos. Although SFT does not require teacher scores, it trains the AR student under teacher-induced prefixes rather than under the student's own rollout distribution. At inference time, the student conditions on

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.

its previously generated frames; local errors can therefore shift future denoising states and accumulate over the video horizon (Bengio et al., 2015). A preliminary SFT sweep in Figure 1 confirms this mismatch: longer off-policy training fails to consistently improve either VBench (Huang et al., 2024) or VideoPhy-2 (Bansal et al., 2026).

The SFT mismatch motivates on-policy adaptation, yet existing on-policy objectives still assume supervision unavailable in black-box video distillation. DMD-style objectives require teacher scores, density ratios, or compatible noised states (Yin et al., 2024), none of which are available from a sampling-only video teacher. Reward- and preference-based diffusion alignment can optimize models from sample-level feedback (Black et al., 2023; Wallace et al., 2024; Liu et al., 2025b;a; Xue et al., 2025), but a scalar score on a completed video does not identify which frames, motion patterns, or denoising states account for the teacher–student discrepancy. The useful signal is observable only at completed videos, while the object being trained is a time-dependent vector field evaluated on the student's own noised AR states. The key challenge is therefore not merely to assign a reward to a video, but to lift black-box video evidence into dense flow-matching supervision for the student's causal denoising process.

We propose *Adversarial Flow Distillation* (AFD), an on-policy distillation framework for autoregressive video generation under heterogeneous black-box teacher access. AFD factorizes the problem into clean-sample distribution-ratio estimation and forward-process vector-field regression. For each prompt batch, it queries the teacher for completed videos, rolls out the current student under autoregressive self-conditioning, and trains a prompt-conditioned discriminator on teacher samples versus current student samples. The discriminator produces an on-policy advantage score, normalized against the current batch or prompt distribution, which is identifiable from black-box samples and co-evolves with the student.

AFD then uses this advantage inside a reward-weighted flow-matching objective rather than a reverse-trajectory policy-gradient objective. Student rollouts are forward-noised by the student's own schedule, producing the same kind of intermediate states on which its velocity field is trained. Within each on-policy batch, high-advantage rollouts form positive examples and low-advantage rollouts form negative examples; the student is then trained to move its vector field toward the forward velocity of higher-scoring rollouts and away from lower-scoring rollouts, with the discriminator advantage controlling the strength of this contrastive correction. This turns a black-box video-level signal into denoising-time updates on the student's own noised states. Consequently, AFD requires no teacher scores, no teacher–student step alignment, and no stored reverse trajectories, while still providing dense supervision beyond video-level adversarial training. The population analysis below makes the role of the score precise: its centered form is a strict local reverse-KL descent direction, and reweighting it changes the student's conditional forward-flow target. Our contributions are as follows:

- We identify black-box heterogeneous on-policy distillation as a core obstacle for autoregressive video students, and show that off-policy SFT, score-based DMD, and direct video-level adversarial training are mismatched to the limited teacher interface.
- We introduce Adversarial Flow Distillation (AFD), a score-free distillation framework that estimates teacher–student discrepancy from completed videos and converts it into dense forward-process flow-matching updates on the student's own noised rollouts.
- We evaluate AFD on two causal autoregressive video backbones, showing gains on motion- and physics-sensitive metrics while preserving general video quality. Ablation studies further show the importance of online discriminator updates and rollout-specific feedback.

2 METHOD

Let $\pi_T(x_0|y)$ denote a black-box teacher that returns a video for prompt y, and let π_θ denote a causal autoregressive student flow model with velocity field $f_\theta(x_t, t, y)$. The teacher may differ from the student in architecture, capacity, latent representation, and sampling schedule. We assume *no access to* teacher parameters, scores, latents, or denoising trajectories; the only shared interface is the completed prompt-conditioned video. At each iteration, prompts are sampled from a distribution $y \sim \mathcal{Y}$, the teacher returns $x_0^T \sim \pi_T(\cdot|y)$, and the current student produces an on-policy rollout $\hat{x}_0 \sim \pi_\theta(\cdot|y)$ under autoregressive self-conditioning. As shown in Figure 2, AFD consists of an

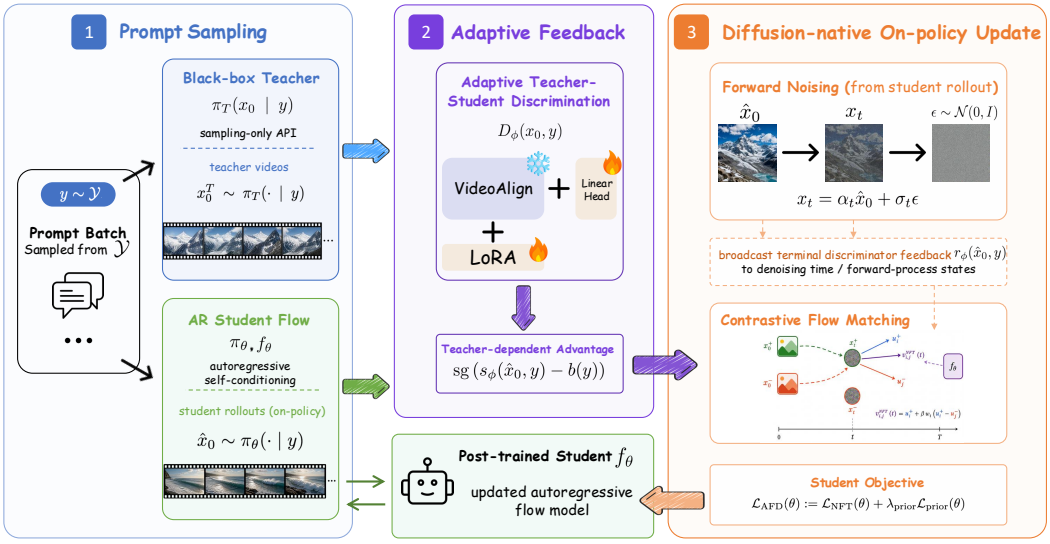

Figure 2: Overview of AFD for black-box on-policy video distillation. A sampling-only teacher provides prompt-conditioned completed videos, while the current AR student provides on-policy rollouts. A co-evolving discriminator estimates teacher–student distributional discrepancy from completed samples, and DiffusionNFT converts this black-box sample-level signal into dense forward-process flow updates on the student's own noised states.

adaptive video discriminator that estimates teacher–student distributional discrepancy on student rollouts and a DiffusionNFT update that transfers this sample-level signal to the student's forward noising process.

2.1 Adaptive Teacher–Student Discrimination

We train a prompt-conditioned spatiotemporal discriminator $D_\phi(x_0, y)$ that scores how likely a video x_0 is to be a teacher sample given prompt y. Define $s_\phi(x_0, y) = \log[D_\phi(x_0, y)/(1 - D_\phi(x_0, y))]$ for its BT logit, as depicted in Figure 2. For each prompt, we treat the teacher sample as preferred over the current student rollout, yielding the Bradley–Terry (BT) loss:

$$\mathcal{L}_D(\phi) := -\mathbb{E}_{(x_0^T, \hat{x}_0, y)} \log \sigma\big(s_\phi(x_0^T, y) - s_\phi(\hat{x}_0, y)\big), \tag{1}$$

where $\sigma(\cdot)$ is the logistic function. This pairwise objective is standard in preference modeling: it increases the discriminator margin between teacher samples and current student samples under the same prompt, without requiring calibrated absolute rewards. In practice, D_ϕ can be initialized from a video preference or text-video alignment model such as VideoAlign (Liu et al., 2025b) and adapted with LoRA. Rather than defining a fixed global reward, the discriminator estimates the current discrepancy between teacher and student samples under the shared prompt distribution, providing on-the-fly feedback for on-policy rollouts when teacher scores are unavailable.

The discriminator induces a discrepancy signal, which we use as an adaptive reward function:

$$r_\phi(\hat{x}_0, y) = \mathrm{sg}\left(s_\phi(\hat{x}_0, y) - b(y)\right), \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $b(y)$ is a batch or prompt-level baseline. Unlike a fixed reward model, this signal co-evolves with the student model and measures whether current student videos remain distinguishable from teacher videos under the current prompt distribution.

2.2 Diffusion-Native On-Policy Update

The discriminator signal is sample-level because it scores completed videos. Directly optimizing this signal with reverse-trajectory policy gradients would require treating the reverse denoising chain as an RL trajectory, which is costly for long, high-resolution videos. We instead adopt a forward-process

diffusion optimization method that optimizes diffusion models using only clean samples and the forward noising process (Choi et al., 2026), namely DiffusionNFT (Zheng et al., 2025). For a student rollout $\hat{x}_0$, timestep $t \in [0,1]$, and noise $\epsilon \sim \mathcal{N}(0, I)$, define the forward-noised sample and its corresponding flow-matching target as

$$x_t = \alpha_t \hat{x}_0 + \sigma_t \epsilon, \qquad v = \dot{\alpha}_t \hat{x}_0 + \dot{\sigma}_t \epsilon. \tag{3}$$

We use the discriminator to score on-policy rollouts with π_θ. We then determine the weights for positive and negative policy optimization. Concretely, for a minibatch of student videos $\{\hat{x}_{0,i}\}_{i=1}^B$, let $A_i = r_\phi(\hat{x}_{0,i}, y_i)$ and normalize it to a weight $w_i \in [0,1]$ within the batch. We apply the forward process on this batch of videos following Equation 3 to obtain noised samples $\{x_{t,i}\}_{i=1}^B$. Denoting the parametrized velocity as $v_\theta(x_{t,i}, t, y_i)$, we define two operators for positive and negative policy optimization under the on-policy sampling setting:

$$v_\theta^+ := (1-\beta)\,\mathrm{sg}(v_\theta) + \beta v_\theta, \qquad v_\theta^- := (1+\beta)\,\mathrm{sg}(v_\theta) - \beta v_\theta.$$

the negative-aware fine-tuning (NFT) loss whose optimum implicitly drives v_θ towards v^+ is

$$\mathcal{L}_{\text{NFT}}(\theta) = \mathbb{E}_{t,\,x_{t,i}}\Big[\sum_{i=1}^{B} w_i \cdot \|v_\theta^+(x_{t,i}, t, y_i) - v\|^2 + (1-w_i)\cdot \|v_\theta^-(x_{t,i}, t, y_i) - v\|^2\Big], \tag{4}$$

Additionally, we include a regularization term in the objective function to avoid catastrophic forgetting $\mathcal{L}_{\text{prior}} = \mathbb{E}_{t,\,x_{t,i}} \sum_{i=1}^B w_i \|v_\theta(x_{t,i}, t, y_i) - v^{\text{ref}}(x_{t,i}, t, y_i)\|^2$, where the reference model v^{ref} is the initial pre-trained diffusion model. The objective of our method AFD is finally defined as:

$$\mathcal{L}_{\text{AFD}}(\theta) := \mathcal{L}_{\text{NFT}}(\theta) + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}(\theta). \tag{5}$$

2.3 Black-Box Teacher Interface

AFD is motivated by the measurability constraint imposed by a black-box video teacher. The teacher exposes only the clean-sample channel

$$\mathcal{O}_T: \quad y \mapsto x_0^T \sim \pi_T(\cdot|y), \tag{6}$$

not $\nabla_x \log p_T(x_t|y)$, teacher latents, teacher reverse states, or a teacher transition kernel. Any admissible distillation signal must therefore be a function of completed teacher videos and completed student videos. This constraint is restrictive for an AR flow student, because the trained object is not a video classifier but a vector field $f_\theta(x_t, t, h_k, y)$ evaluated at noised states along student-induced histories $h_k = x_0^{(<k)}$.

Let $x_0 = [x_0^{(1)}, \ldots, x_0^{(K)}]$ denote a sequence of video blocks. The AR student induces

$$\pi_\theta(x_0|y) = \prod_{k=1}^{K} \pi_\theta\Big(x_0^{(k)} | x_0^{(<k)}, y\Big), \tag{7}$$

so student errors alter the future conditioning distribution. A bidirectional black-box teacher, however, is observed only through $\pi_T(x_0|y)$. Its hidden sampler may contain M_T denoising evaluations, whereas the student may use $M_S \ll M_T$. Without teacher trajectories or a shared latent space, there is no observable alignment operator

$$\mathcal{A}_{m,s}: z_m^T \mapsto (h_k, x_s^S) \tag{8}$$

from a teacher step m to a student AR denoising state s. The absence of such an operator clarifies the limitations of two common baselines. SFT trains under teacher-induced prefixes rather than the student's rollout distribution, so exposure bias accumulates along the video horizon (Bengio et al., 2015). DMD-style objectives require either a teacher score or a teacher density ratio at the student's noised state,

$$\nabla_x \log p_T(x_t|y) \quad \text{or} \quad \log \frac{\pi_T(\hat{x}_0|y)}{\pi_\theta(\hat{x}_0|y)}. \tag{9}$$

The score term is unobservable, and the clean-sample ratio can only be estimated from samples at x_0. The relevant objective is therefore not to reconstruct the teacher's hidden diffusion path, but to

transfer completed-video evidence to the student's own noised states. DiffusionNFT matches this interface: it requires only clean samples from the current policy, a scalar score on those samples, and the student's known forward noising kernel.

This differs from GRPO-style diffusion RL. Methods such as Flow-GRPO (Liu et al., 2025a) and DanceGRPO (Xue et al., 2025) make diffusion or flow policies amenable to online RL by casting the reverse denoising process as an MDP, introducing stochastic reverse trajectories, and optimizing group-relative policy objectives on sampled rollouts. This interface is well suited to reward-model alignment, where the optimized model's reverse sampler defines the environment. It is less suitable for black-box teacher distillation: the teacher provides no reverse actions, no teacher trajectory probabilities, and no step-level supervision on the student's reverse chain. Applying a reverse-MDP objective therefore introduces a synthetic credit-assignment layer that is not supported by the teacher interface. AFD instead keeps the teacher-dependent signal at the clean-sample level, where it is identifiable, and uses the forward noising kernel to induce the dense denoising-time structure.

2.4 Teacher-Directed Population Update and Forward-Process Propagation

Fix a prompt y and write $P_y = \pi_T(\cdot \mid y)$ and $Q_y = \pi_\theta(\cdot \mid y)$ for teacher videos and current student rollouts. Assume that they have common support and let $\rho_y(x_0) = \log(P_y(x_0)/Q_y(x_0))$. The population Bradley–Terry comparison identifies this density ratio up to a prompt-dependent constant: $s^*(x_0, y) = \rho_y(x_0) + C(y)$. With $b(y) = \mathbb{E}_{Q_y} s^*(x_0, y)$, the centered score in Eq. 2 is precisely $\rho_y - \mathbb{E}_{Q_y}\rho_y$.

Proposition 1 (Exact teacher tilt and local reverse-KL descent)**.** *Suppose the BT scorer is population-optimal and unrestricted. Then the normalized exponential tilt $Q_y^{\exp}(x_0) \propto Q_y(x_0) \exp s^*(x_0, y)$ equals P_y. If ρ_y is bounded and $\mathrm{Var}_{Q_y}(\rho_y) > 0$, the normalized linear tilt $Q_y^\eta = g_\eta Q_y$, with $g_\eta = 1 + \eta(\rho_y - \mathbb{E}_{Q_y}\rho_y)$, satisfies $\mathrm{KL}(Q_y^\eta \| P_y) < \mathrm{KL}(Q_y \| P_y)$ for every sufficiently small $\eta > 0$.*

The exact recovery statement follows by cancelling Q_y against $\exp \rho_y = P_y/Q_y$. The strict descent statement follows from an identity valid for any $g \geq 0$ with $\mathbb{E}_{Q_y} g = 1$:

$$\mathrm{KL}(gQ_y \| P_y) - \mathrm{KL}(Q_y \| P_y) = \mathrm{KL}(gQ_y \| Q_y) - \mathrm{Cov}_{Q_y}(g, \rho_y). \tag{10}$$

For g_η, the covariance is $\eta \,\mathrm{Var}_{Q_y}(\rho_y)$, whereas the first term is $O(\eta^2)$. Hence

$$\mathrm{KL}(Q_y^\eta \| P_y) - \mathrm{KL}(Q_y \| P_y) = -\eta \,\mathrm{Var}_{Q_y}(\rho_y) + O(\eta^2) < 0 \tag{11}$$

for sufficiently small η. Boundedness ensures $g_\eta \geq 0$ in this range. Proposition 1 thus supplies a strict improvement result for the centered teacher–student signal, while the full exponential tilt identifies its exact population target. The proof is in Appendix C.

The signal also reaches the student's noised states without a teacher trajectory. Let $u_t = \dot{\alpha}_t x_0 + \dot{\sigma}_t \epsilon$ be the forward-path velocity in Eq. 3. For any nonnegative clean-video weight $g(x_0)$, the squared-error flow-matching minimizer under the reweighted student law obeys

$$v_g^*(x_t, t, y) - v_0^*(x_t, t, y) = \frac{\mathrm{Cov}_{Q_y}(g(x_0), u_t \mid x_t, t, y)}{\mathbb{E}_{Q_y}[g(x_0) \mid x_t, t, y]}. \tag{12}$$

The posterior over clean rollouts varies with x_t and t, so a scalar assigned to each completed video changes the conditional vector-field target across noised states. This is rollout-conditioned propagation, rather than a claim that the scalar identifies the responsible frame. Appendix C derives Eq. 12. The within-prompt reward-shuffling ablation in Table 6 tests the importance of keeping each score paired with its generating rollout.

2.5 Training Procedure

The optimization alternates between on-policy data collection, discriminator updates, and student velocity-field updates, as summarized in Algorithm 1. We implement the prior regularizer $\mathcal{L}_{\mathrm{prior}}(\theta)$ as a velocity-regression penalty against a frozen reference student to preserve the student's base capabilities.

Algorithm 1: Adversarial Flow Distillation (AFD)

Input: Teacher API π_T, Student Flow Model f_θ, Discriminator D_ϕ, Prompt distribution $\mathcal{Y}$
Output: Post-trained Student Model f_θ
Initialize student parameters θ and EMA parameters $\bar{\theta} \leftarrow \theta$, and keep f_{ref} frozen
Initialize discriminator parameters ϕ (e.g., via LoRA on VideoAlign)
while *not converged* **do**
  `// 1. On-Policy Data Collection`
  Sample batch of prompts $y \sim \mathcal{Y}$
  Query teacher for target videos: $x_0^T \sim \pi_T(\cdot|y)$
  Sample student rollouts: $\hat{x}_0 \sim \pi_\theta(\cdot|y)$
  `// 2. Adaptive Teacher-Student Discrimination`
  Compute discriminator loss $\mathcal{L}_D(\phi)$ comparing x_0^T and $\hat{x}_0$
  Update discriminator $\phi \leftarrow \phi - \eta_D \nabla_\phi \mathcal{L}_D(\phi)$
  Compute baseline $b(y)$ (e.g., batch mean of $s_\phi(\hat{x}_0, y)$)
  Compute advantage scores $r_\phi(\hat{x}_0, y) = \text{sg}(s_\phi(\hat{x}_0, y) - b(y))$
  `// 3. Diffusion-Native Student Update`
  Sample denoising timesteps $t \sim \mathcal{U}[0, 1]$ and noise $\epsilon \sim \mathcal{N}(0, I)$
  Construct forward-noised student states $x_t = \alpha_t \hat{x}_0 + \sigma_t \epsilon$
  Form positive/negative rollout sets from r_ϕ and evaluate $\mathcal{L}_{\text{NFT}}(\theta)$ using x_t
  Evaluate prior regularization $\mathcal{L}_{\text{prior}}(\theta)$ against reference student f_{ref}
  Update student $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \left(\mathcal{L}_{\text{NFT}} + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}\right)$
  `// 4. EMA Update`
  Update target network $\bar{\theta} \leftarrow \beta \bar{\theta} + (1 - \beta)\theta$

3 Experiments

3.1 Setup

Experimental setup. We evaluate AFD on two causal autoregressive student families, Self-Forcing and Causal-Forcing. The teacher is Seedance 2.0 (ByteDance Seed, 2026), accessed only as a prompt-conditioned video sampling API. The main experiment is a continual adaptation setting: a pretrained AR student is adapted to a physics-oriented target domain while preserving general video quality. We sample 200 examples from the VideoPhy-2 physics benchmark and query the teacher on the same prompts to obtain black-box adaptation videos. Unless otherwise specified, discriminators are initialized from VideoAlign (Liu et al., 2025b), adapted with LoRA, and updated online against the current student. We report VBench (Huang et al., 2024) dimensions grouped into Physics and General categories. Physics includes temporal flickering, motion smoothness, dynamic degree, human action, and spatial relationship; General is the mean over the remaining VBench dimensions. We also evaluate target-domain adaptation with VideoAlign Motion Quality (VideoAlign-MQ) (Liu et al., 2025b) and VideoPhy-2 Physical Consistency (VideoPhy-2-PC) (Bansal et al., 2026). Full hyperparameters are provided in Appendix A.

Baselines. We consider four baselines:

- **Base**: the pretrained AR student before teacher adaptation.
- **SFT**: supervised fine-tuning on teacher-generated videos without on-policy student rollouts.
- **GAN**: adversarial video-level training with the teacher–student discriminator, excluding the forward-process policy update.
- **GAN + diffusion-loss**: GAN training supplemented by a student-side diffusion loss, without DMD's teacher-score distribution-matching term.

3.2 Main Results

Tables 1 and 2 show that AFD improves physics-sensitive generation under limited target-domain data while preserving general video generation capability. On Self-Forcing, AFD reaches the best Physics VBench Total (87.55), VideoAlign-MQ (0.605), and VideoPhy-2-PC (4.20), with a General

Total close to the best baseline (60.83 vs. 60.95). On Causal-Forcing, AFD gives the best Physics VBench Total (88.52) and VideoPhy-2-PC (4.24), ties the best VideoAlign-MQ (0.661), and keeps the General Total close to the highest score (59.83 vs. 60.32). Compared with the adapted baselines, AFD gives the strongest dynamic degree on both student families, indicating better motion and physical behavior without a large drop in general quality.

Table 1: Main VBench results after black-box continual adaptation. Physics reports all dimensions and their average; General reports the group average.

Model	Method	Physics (VBench)						General (VBench)
		Total↑	Temporal Flick. ↑	Motion Smooth. ↑	Dynamic Degree ↑	Human Action ↑	Spatial Rel. ↑	Total↑
Self-Forcing	Base	68.49	98.27	97.19	91.67	46.00	9.32	36.51
	SFT	58.69	96.64	97.40	44.44	34.00	20.97	37.62
	GAN	83.83	99.28	98.58	61.11	79.00	81.18	60.95
	GAN + diff-loss	81.88	99.01	98.95	52.78	81.00	77.66	59.98
	AFD	**87.55**	98.78	98.30	79.17	80.00	81.50	60.83
Causal-Forcing	Base	86.76	98.20	97.90	87.50	76.00	74.20	59.03
	SFT	76.24	99.31	99.05	29.17	78.00	75.69	59.40
	GAN	83.07	98.58	98.08	66.67	76.00	76.00	59.44
	GAN + diff-loss	82.52	99.16	98.75	58.33	77.00	79.35	60.32
	AFD	**88.52**	98.41	97.89	88.89	80.00	77.39	59.83

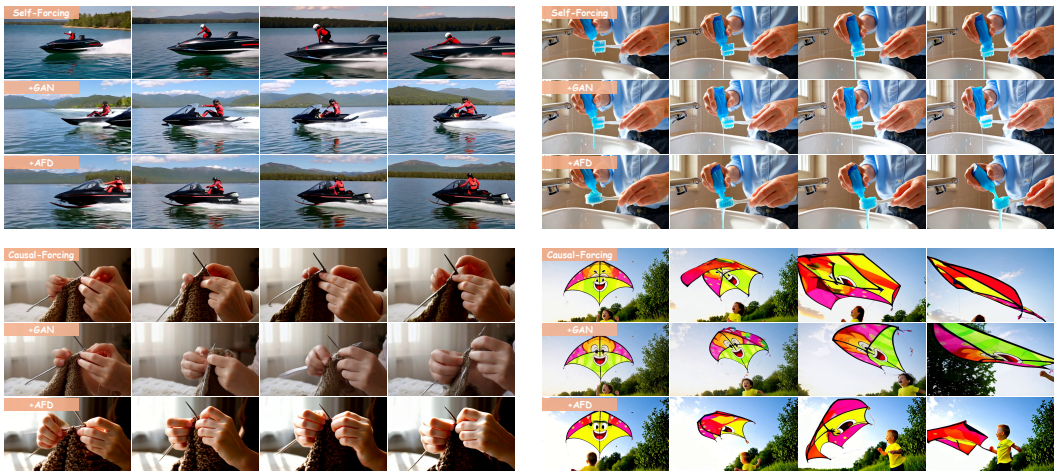

Figure 3: Qualitative visualizations of AFD in the main experiment. The first row shows Self-Forcing examples, and the second row shows Causal-Forcing examples. Each example shows representative teacher–student comparison frames under the same prompt, illustrating how the on-policy AFD update changes the student's generated video distribution. Prompts are given in Table 8.

Independent physics diagnostic. We evaluate the same Self-Forcing videos with a frozen VisionReward-Video judge (Xu et al., 2026). Its score averages three physics-law Yes/No answers. As shown in Table 3, AFD scores 0.973, compared with 0.923 for GAN and 0.907 for the base student. The 0.050 margin over GAN, the strongest baseline in this diagnostic, is consistent with AFD's lead on Physics VBench and VideoPhy-2-PC in Tables 1 and 2. Because neither the judge nor its fixed scoring rule is used in training, this agreement provides complementary evidence for the observed physics gains

Table 2: Physics evaluations after small-sample physics-domain adaptation.

Model	Method	Motion Quality↑	Physical Consistency↑
Self-Forcing	Base	0.341	4.04
	SFT	0.296	3.72
	GAN	0.541	4.10
	GAN + diff-loss	0.420	3.72
	AFD	**0.605**	**4.20**
Causal-Forcing	Base	0.520	4.16
	SFT	0.499	4.17
	GAN	0.582	4.16
	GAN + diff-loss	0.661	4.14
	AFD	**0.661**	**4.24**

Table 3: Independent VisionReward-Physics diagnostic on Self-Forcing videos. The score averages three physics-law Yes/No answers from a frozen judge.

Method	VisionReward-Physics↑
Base	0.907
SFT	0.891
GAN	0.923
GAN + diff-loss	0.919
AFD	**0.973**

Figure 3 illustrates the motion and physical-consistency trends in Tables 1 and 2 across both AR student families. Further examples are in Appendix D.

Teacher and architecture robustness. Table 4 tests whether AFD's gains persist across sampling-only teachers and student architectures. With Self-Forcing, Seedance 1.5 and Kling 2.5 positives raise Physics VBench from 68.49 for the base student to 85.50 and 85.70, respectively. Their similar scores suggest that the gain does not depend on the choice between these two positive sources. On the bidirectional Wan2.1 student, the same teachers raise Physics VBench from 84.02 to 85.72 and 85.69. The absolute gain is smaller from Wan2.1's higher baseline, but its consistency across teachers shows that the improvement also extends to the tested bidirectional student.

Table 4: Robustness to alternative sampling-only teachers and the Wan2.1 student. Each adapted row uses cached positives from the listed source.

Setting	Positive source	Physics VBench↑	General VBench↑
Self-Forcing base	–	68.49	36.51
Self-Forcing AFD	Seedance 1.5	85.50	61.14
Self-Forcing AFD	Kling 2.5	85.70	61.05
Wan2.1 base	–	84.02	56.30
Wan2.1 AFD	Kling 2.5	85.69	57.11
Wan2.1 AFD	Seedance 1.5	85.72	57.01

AFD also improves Self-Forcing with real-video positives from OpenVid (Nan et al., 2025). Even with sampling-only access, the teacher provides prompt-matched positives from one generator with standardized outputs. Across equal-budget draws, these videos yield higher mean performance and lower variation, consistent with a more coherent target for on-policy distillation (Appendix Table 10).

3.3 Ablations

Discriminator learning rate. Because AFD uses a co-evolving discriminator, its update rate η_D affects the feedback available to the student throughout training. We sweep $\eta_D \in \{0, 1\times10^{-6}, 5\times10^{-6}, 1\times10^{-5}, 5\times10^{-5}\}$ with other settings fixed and track the discriminator output $D_\phi(\hat{x}_0, y)$ on student rollouts. Figure 4 shows that the output approaches one at $\eta_D = 1\times10^{-6}$, where the discriminator lags behind the changing student. At 5×10^{-5}, it stays near zero as the discriminator becomes too strong. The intermediate rates avoid both saturation regimes and produce a gradual increase in the output.

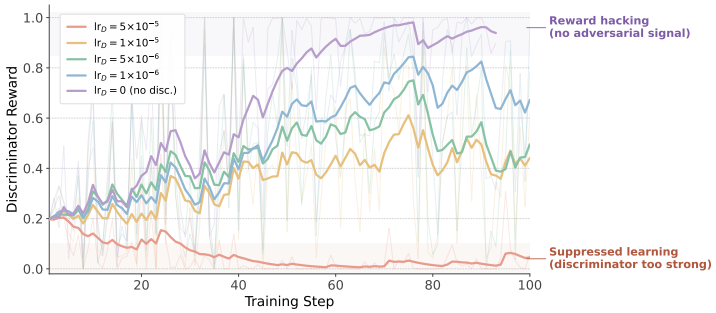

Figure 4: Discriminator learning-rate ablation. Intermediate update rates yield a gradual increase in discriminator output on student rollouts, while extreme rates lead to saturation.

Online discrimination. Table 5 compares updated and frozen discriminators using the same Kling 2.5 positives. Updating the discriminator at $\eta_D = 5\times10^{-6}$ raises Physics VBench from 84.42 to 85.70 and Dynamic Degree from 69.44 to 73.61. These gains show the value of adapting the feedback to the changing student under the tested settings.

Table 5: Effect of online discriminator updates using the same Kling 2.5 positives.

Setting	Physics VBench↑	Dynamic Degree↑	General VBench↑	VideoPhy-2 PC↑
Frozen ($\eta_D = 0$)	84.42	69.44	61.48	4.20
Online ($\eta_D = 5\times10^{-6}$)	85.70	73.61	61.05	4.22

Rollout-specific feedback. To examine the role of rollout-specific feedback, we shuffle reward values among rollouts of the same prompt while preserving the reward multiset. Physics VBench falls from 85.38 with correct pairing to 81.95 after shuffling (Table 6). Because the prompts and reward values remain fixed, the drop supports the importance of pairing each score with its generating rollout.

Table 6: Effect of rollout–reward pairing using Kling 2.5 positives. Shuffling preserves the reward multiset within each prompt.

Setting	Physics VBench↑	General VBench↑	VideoPhy-2-PC↑
Correct pairing	85.38	60.99	4.18
Rewards shuffled within prompt	81.95	60.17	4.04

We further examine the role of adaptive, rollout-specific feedback by varying the discriminator update rate, replacing the online discriminator with a frozen VideoAlign reward, and assigning uniform rollout weights (Appendix Table 11). We also assess how the training objective and discriminator initialization affect performance (Appendix Table 12).

4 Conclusion

AFD suggests that black-box video distillation should be formulated around the information interface available to the student, rather than around the teacher's hidden generation process. For causal AR students, the missing ingredient is not access to the teacher trajectory, but a principled way to connect observable sample-level discrepancies to the student's own denoising-time states. By separating these two roles into adaptive discrimination and forward-process credit assignment, AFD turns completed-video supervision into an on-policy training signal that improves motion and physical behavior without sacrificing general quality.

AI USE STATEMENT

We used generative AI tools to help refine the presentation of mathematical arguments and polish the analysis of reported experimental results. We did not use generative AI tools to generate synthetic datasets, implement the method, or process experimental data. Additionally, we used generative AI tools to draft and polish selected passages and assist with literature searches and reference checks. The authors reviewed and revised all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *Proceedings of the International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in Neural Information Processing Systems*, 28:1171–1179, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. URL https://arxiv.org/abs/2305.13301.

Tim Brooks, Bill Peebles, et al. Video generation models as world simulators. OpenAI technical report, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

ByteDance Seed. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. URL https://arxiv.org/abs/2604.14148.

Ethan Chern, Zhulin Hu, Bohao Tang, Jiadi Su, Steffi Chern, Zhijie Deng, and Pengfei Liu. LiveTalk: Real-time multimodal interactive video diffusion via improved on-policy distillation. *arXiv preprint arXiv:2512.23576*, 2025. URL https://arxiv.org/abs/2512.23576.

Jaemoo Choi, Yuchen Zhu, Wei Guo, Petr Molodyk, Bo Yuan, Jinbin Bai, Yi Xin, Molei Tao, and Yongxin Chen. Rethinking the design space of reinforcement learning for diffusion models: On the importance of likelihood estimation beyond loss design. *arXiv preprint arXiv:2602.04663*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://vchitect.github.io/VBench-project/.

Jisoo Kim, Wooseok Seo, Junwan Kim, Seungho Park, Sooyeon Park, and Youngjae Yu. V.I.P.: Iterative online preference distillation for efficient video diffusion models. *arXiv preprint arXiv:2508.03254*, 2025. URL https://arxiv.org/abs/2508.03254.

Kling Team. Kling-Omni technical report. arXiv preprint arXiv:2512.16776, 2025. URL https://arxiv.org/abs/2512.16776.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025a. URL https://arxiv.org/abs/2505.05470.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025b. URL https://arxiv.org/abs/2501.13918.

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. URL https://arxiv.org/abs/2209.03003.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025. URL https://arxiv.org/abs/2512.04678.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0396ca5a4c628936609aa819bfbca916-Abstract-Conference.html.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, and Zijian He. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. URL https://arxiv.org/abs/2410.13720.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proceedings of the European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/11557_ECCV_2024_paper.php.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Jiazheng Xu, Yu Huang, Jiale Cheng, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. doi: 10.1609/aaai.v40i13.38107. URL https://ojs.aaai.org/index.php/AAAI/article/view/38107.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. URL https://arxiv.org/abs/2505.07818.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025. URL https://arxiv.org/abs/2511.10643.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.18828.

Songchun Zhang, Zeyue Xue, Siming Fu, Jie Huang, Xianghao Kong, Yue Ma, Haoyang Huang, Nan Duan, and Anyi Rao. Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. *arXiv preprint arXiv:2603.17051*, 2026. URL https://arxiv.org/abs/2603.17051.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online Diffusion Reinforcement with Forward Process. *arXiv preprint arXiv:2509.16117*, 2025. URL https://arxiv.org/abs/2509.16117.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62307–62331, 2024. URL https://proceedings.mlr.press/v235/zhou24x.html.

A HYPERPARAMETERS

We provide detailed hyperparameters in Table 7. The table lists the model, optimization, diffusion, regularization, and rollout settings used for AFD training.

Table 7: Comprehensive hyperparameters for AFD training. We detail the configurations used across model and video specifications, LoRA fine-tuning, optimization, diffusion process, selective regularization, and streaming rollout.

Module	Hyperparameter	Value
Model & Video Specs	Base Architecture	Self-Forcing / Causal-Forcing
	Video Resolution ($H \times W$)	480×832
LoRA Fine-Tuning	Rank (r)	256
	Scaling Factor (α)	256
	Dropout Rate	0.0
	Gradient Checkpointing	Enabled
Optimization	Precision Mode	bfloat16
	Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$)
	Learning Rate (η)	1e-5
	Weight Decay	1e-4
	Max Gradient Norm	1.0
Selective Regularization	Interpolation Strength (β)	0.1
	Prior Loss Weight (λ_{prior})	1e-4
	Advantage Clip Max	5.0
	Reward Normalization	Global Std with Per-Prompt Tracking
	EMA Decay Rate (γ)	0.99

B RELATED WORK

Video diffusion and flow models. Denoising diffusion models learn from noised states (Ho et al., 2020), as do score-based models (Song et al., 2021). Flow Matching (Lipman et al., 2023) and Rectified Flow (Liu et al., 2022) instead train continuous-time vector fields. Their acceleration methods differ in what they require from a teacher. Consistency distillation queries a pretrained score model along its probability-flow ODE trajectory. Consistency training can instead learn from data, but does not distill a sampling-only teacher (Song et al., 2023). SiD (Zhou et al., 2024) and DMD (Yin et al., 2024) use teacher scores or denoising predictions at noised generator samples, which a prompt-to-video API does not expose.

Black-box on-policy distillation. On-policy distillation trains on student-generated outputs to reduce the mismatch between training and inference (Lu & Thinking Machines Lab, 2025). Black-box language-model distillation also uses student-generated sequences (Ye et al., 2025). GKD obtains teacher token distributions on such sequences (Agarwal et al., 2024). DAgger instead obtains expert actions at learner-visited states (Ross et al., 2011). A sampling-only video teacher provides neither aligned actions on the student's autoregressive prefixes nor velocity targets at its noised states. AFD uses completed teacher samples and current student rollouts to estimate their discrepancy without querying the teacher on student states.

Adversarial and preference-guided alignment. ADD is a direct antecedent to combining adversarial and diffusion supervision. Its discriminator compares generated images with real data, but its distillation loss requires denoised predictions from an accessible teacher (Sauer et al., 2024). Our GAN baseline retains the sample-based adversarial component, whereas the full ADD objective is unavailable here. DiffusionNFT (Zheng et al., 2025) and Astrolabe (Zhang et al., 2026) update models through the forward process using externally specified rewards. Reward Forcing (Lu et al., 2025) weights DMD gradients with a prespecified VLM motion reward, while still requiring scores

at noised samples. AFD instead learns an adaptive teacher–student signal from completed videos and applies it through the student's forward process.

C Population Analysis

Proof of Proposition 1. Fix y and suppress it from the notation. The two possible assignments of a teacher sample and a student sample to an unordered comparison pair (x, z) have likelihoods $P(x)Q(z)$ and $Q(x)P(z)$. Their log odds are

$$\log \frac{P(x)Q(z)}{Q(x)P(z)} = \rho(x) - \rho(z), \qquad \rho = \log(P/Q).$$

An unrestricted BT score reproduces these odds when $s^*(x) - s^*(z) = \rho(x) - \rho(z)$, giving $s^* = \rho + C(y)$. Because $\mathbb{E}_Q e^{\rho} = \int P = 1$, the normalized exponential tilt is $Q^{\exp} = Qe^{s^*}/\mathbb{E}_Q e^{s^*} = P$.

For any $g \geq 0$ such that $\mathbb{E}_Q g = 1$, let $Q^g = gQ$. Direct substitution gives

$$\mathrm{KL}(Q^g \| P) = \mathbb{E}_Q[g \log g] - \mathbb{E}_Q[g\rho], \qquad \mathrm{KL}(Q\|P) = -\mathbb{E}_Q \rho.$$

Subtracting proves Eq. 10. Now write $h = \rho - \mathbb{E}_Q \rho$ and $g_\eta = 1 + \eta h$. Since ρ is bounded, $g_\eta \geq 0$ for sufficiently small positive η, and $\mathbb{E}_Q g_\eta = 1$. Moreover, $\mathrm{Cov}_Q(g_\eta, \rho) = \eta \,\mathrm{Var}_Q(\rho)$ and $\mathrm{KL}(Q^{g_\eta}\|Q) = \mathbb{E}_Q[(1+\eta h)\log(1+\eta h)] = \frac{1}{2}\eta^2 \,\mathrm{Var}_Q(\rho) + O(\eta^3)$. The strictly negative first-order term in Eq. 11 proves the stated KL descent.

Finite groups and score error. For N student rollouts of the same prompt, normalized exponential group weights based on estimated scores $\hat{\rho}_i$ satisfy

$$\frac{Ne^{\eta\hat{\rho}_i}}{\sum_{j=1}^N e^{\eta\hat{\rho}_j}} = 1 + \eta(\hat{\rho}_i - \overline{\hat{\rho}}) + O(\eta^2). \tag{13}$$

The centered BT score is therefore the first-order group direction of the exponential tilt. At full tilt strength, score errors $\hat{\rho}_i = \rho_i + \delta_i$ change the relative weight of samples i and j by $e^{\delta_i - \delta_j}$. Table 12 compares the exponential and bounded objectives empirically. The proposition makes no finite-score-error assumption.

Derivation of the conditional flow target. Under the student's clean-video law Q_y and forward kernel $p_t(x_t \mid x_0)$, a nonnegative weight $g(x_0)$ changes the posterior over clean videos at state (x_t, t, y) in proportion to $g(x_0)Q_y(x_0)p_t(x_t \mid x_0)$. The squared-error flow-matching minimizer is its conditional velocity mean,

$$v_g^*(x_t, t, y) = \frac{\mathbb{E}_{Q_y}[g(x_0)u_t \mid x_t, t, y]}{\mathbb{E}_{Q_y}[g(x_0) \mid x_t, t, y]}.$$

Subtracting $v_0^* = \mathbb{E}_{Q_y}[u_t \mid x_t, t, y]$ yields Eq. 12, componentwise for vector-valued u_t. The conditional covariance may vary over x_t and t although g is assigned once per complete rollout.

D Additional Qualitative Visualizations

We provide additional AFD visualizations in Figures 5–7. These examples further illustrate the same pattern observed in the main experiments: AFD improves motion evolution and physical plausibility while preserving the prompt-level semantics and visual fidelity of the base autoregressive video student.



Figure 5: Additional qualitative visualization of AFD. Prompt: A breathtaking tightrope walker, clad in a sleek black outfit and a striking red helmet, navigates a narrow tightrope suspended high above a lush green landscape. The camera captures a close-up of their focused face, framed by a dark visor, as they expertly manipulate a wooden block along the rope using a long, slender balance pole. The block glides smoothly, creating a subtle groove in the taut rope, a testament to their precision and control. The background blurs into a vibrant tapestry of greenery, enhancing the sense of isolation and danger. The tightrope, a thin line of tension, contrasts sharply with the serene environment, while the camera remains steady, allowing the viewer to fully immerse in the walker's intense concentration and the delicate dance of balance and skill.



Figure 6: Additional qualitative visualization of AFD. Prompt: A majestic dragon kite soars through a narrow urban canyon, framed by towering, weathered brick buildings. The kite, intricately crafted with a vibrant red body and a fierce, golden head, glides effortlessly against a backdrop of deep blue skies, its tail adorned with a cascade of colorful ribbons that flutter in the breeze. The camera captures the kite's dynamic flight, showcasing its graceful turns and the subtle interplay of light and shadow as it navigates the tight space between the buildings. The scene is enhanced by a soft, golden-hour glow, casting a warm hue over the textured brick facades, while the kite's movement creates a striking contrast against the static architecture. This cinematic moment celebrates the artistry of kite flying, blending the beauty of nature with the urban landscape.

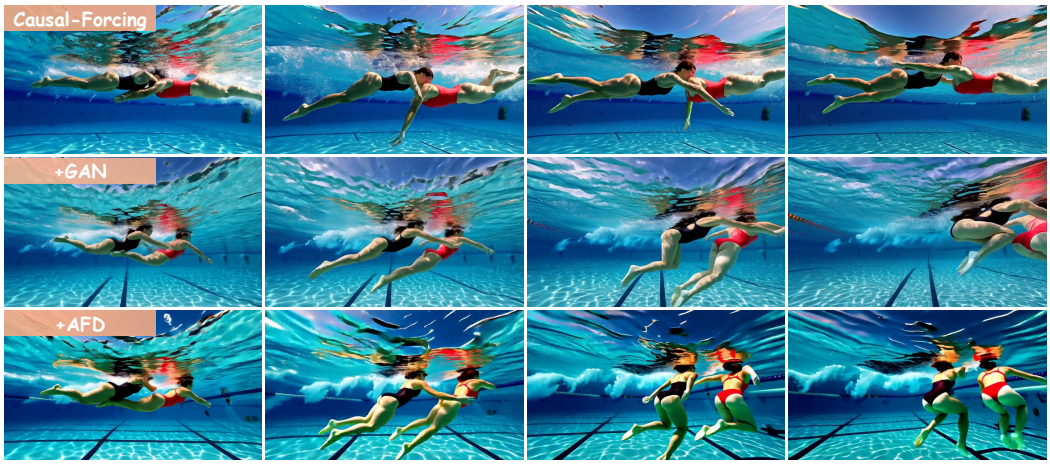

Figure 7: Additional qualitative visualization of AFD. Prompt: In a breathtaking underwater tableau, two swimmers glide through the crystal-clear water, their synchronized breaststroke movements creating a mesmerizing dance of rhythm and grace. The camera, positioned at a slight angle, captures the swimmers' powerful legs and arms as they propel themselves forward, their bodies slicing through the water with precision. The first swimmer, clad in a sleek black swimsuit, leads the pack, while the second, in a vibrant red suit, trails closely, their synchronized movements echoing a fierce competition. The water's surface shimmers with sunlight, casting dynamic reflections that enhance the scene's ethereal quality. The static camera allows for a focused view of their synchronized strokes, highlighting the swimmers' strength and agility, while the serene underwater environment amplifies the beauty of their synchronized performance.

E PROMPTS FOR QUALITATIVE VISUALIZATIONS

We present the prompts used for the qualitative examples in Figures 3 and 8.

Table 8: Prompts used for Figure 3.

Example	Prompt
physics1	In a breathtaking display of aquatic athleticism, a skilled skier glides effortlessly across the shimmering surface of a tranquil lake, propelled by a powerful ski jet. The scene opens with the skier, clad in a vibrant red and black wetsuit, gracefully poised on the water, their skis slicing through the glassy surface. The ski jet, a sleek black vessel with a striking red stripe, glides alongside, its engine humming with precision as it maintains a steady pace. The camera captures the dynamic interplay between the skier and the jet, showcasing the skier's agility and balance as they lean into the turns, their movements fluid and controlled. The backdrop features a serene landscape of lush greenery and distant mountains, enhancing the tranquil ambiance. As the skier gains momentum, the camera remains fixed, allowing viewers to appreciate the exhilarating dance of speed and skill on the water, culminating in a thrilling display of aquatic prowess.
physics2	In a serene bathroom setting, a person clad in a light blue shirt and dark pants stands before a sleek white sink, the soft glow of natural light filtering through a nearby window. The camera captures a close-up of their hands, one gently squeezing a vibrant blue toothpaste tube, while the other holds a pristine white toothbrush, poised to receive the creamy substance. As the tube is squeezed, a steady stream of toothpaste flows onto the bristles, creating a mesmerizing cascade of blue that contrasts beautifully with the white brush. The scene is bathed in a warm, inviting light, enhancing the textures of the toothpaste and the brush, while the background remains softly blurred, drawing the viewer's focus to the meticulous act of oral hygiene. The static camera angle ensures a clear view of the process, capturing the rhythmic motion of the hands and the satisfying result of a perfectly coated toothbrush.
physics3	In a serene, softly lit room, a pair of hands gracefully manipulates knitting needles, their rhythmic movements creating a mesmerizing dance of craftsmanship. The camera captures a close-up of the hands, framed against a blurred backdrop, enhancing the focus on the intricate process. The needles glisten under the warm, diffused light, casting gentle shadows that accentuate the texture of the yarn. As the hands deftly weave the yarn, the knitted piece grows visibly longer, each stitch a testament to the skill and patience of the unseen artisan. The static camera angle allows for an intimate exploration of the technique, while the subtle color grading enhances the cozy atmosphere, inviting viewers to immerse themselves in the artistry of knitting.
physics4	A vibrant kite, adorned with whimsical cartoon characters, soars against a backdrop of lush greenery and a clear blue sky, its colorful design capturing the essence of childhood joy. The kite, featuring a playful character with a wide smile and a distinctive hat, is held by a young child in a bright yellow shirt, who eagerly runs in a circular motion, their laughter echoing through the air. The camera, positioned at a low angle, captures the dynamic interplay between the child's enthusiasm and the kite's graceful ascent, as it dances on the breeze. The scene is bathed in soft, natural light, enhancing the vivid hues of the kite and the child's clothing, while the gentle rustle of leaves and the distant hum of nature create an immersive atmosphere. This moment encapsulates the simple pleasures of outdoor play, inviting viewers to share in the exhilaration of flight and the boundless energy of youth.r

Table 9: Prompts used for Figure 8.

Example	Prompt
anime1	A brave cartoon rabbit exploring a glowing mushroom forest at dusk, expressive 3D cartoon character, magical storybook atmosphere, warm cinematic lighting, whimsical fantasy environment, soft polished textures, original fairy-tale-inspired animation style.
anime2	A cartoon whale floating through a dreamy sky filled with stars and clouds, whimsical fantasy sequence, polished 3D cartoon family-animation look, cinematic soft glow, original fairy-tale-inspired animation style.

F Additional Experimental Results

We examine the effects of positive-video source, feedback design, objective choice, and discriminator initialization. We also provide stylized-domain examples and compare pairwise with binary discriminator supervision.

Positive-source variation. Table 10 compares three equal-budget draws from OpenVid with prompt-matched Seedance 2.0 samples. OpenVid positives achieve 85.12 ± 1.39 Physics VBench, showing that AFD can also learn from real videos. Across these draws, Seedance 2.0 yields higher means on all four metrics and less variation, including 87.28 ± 0.32 Physics VBench. This pattern is consistent with teacher-generated videos providing a more coherent positive target through matched prompts and standardized outputs. The three draws do not establish that teacher samples are preferable to all curated real-video sources.

Table 10: Positive-source comparison for Self-Forcing AFD over three equal-budget draws (mean $\pm$ standard deviation).

Positive source	Physics VBench↑	General VBench↑	VideoAlign-MQ↑	VideoPhy-2-PC↑
OpenVid	85.12 ± 1.39	59.76 ± 2.30	0.514 ± 0.075	4.17 ± 0.03
Seedance 2.0	87.28 ± 0.32	60.48 ± 0.33	0.606 ± 0.005	4.21 ± 0.01

Additional feedback ablations. Table 11 further tests whether the gain comes from adaptive, rollout-dependent feedback. The online discriminator at $\eta_D = 10^{-5}$ reaches 85.38 Physics VBench and 72.22 Dynamic Degree, compared with 83.20 and 70.22 for a frozen VideoAlign reward. Keeping discriminator training but assigning every rollout the same weight reduces Physics VBench to 81.74. Thus, the student-side update alone does not account for the gain. Its effectiveness depends on feedback that distinguishes individual rollouts.

Table 11: Additional feedback ablations on Kling 2.5 positives.

Setting	Physics VBench↑	Dynamic Degree↑	General VBench↑	VideoPhy-2-PC↑
Online discriminator, $\eta_D = 10^{-5}$	85.38	72.22	60.99	4.18
Frozen VideoAlign reward	83.20	70.22	61.20	4.16
Uniform $w_i = 0.5$	81.74	–	60.27	4.12

Training objective and discriminator initialization. Table 12 first compares the bounded two-branch objective with an exponential-tilt objective based on the BT scores. The bounded objective improves Physics VBench from 81.80 to 85.38, with smaller gains in General VBench and VideoPhy-2-PC. The exponential-tilt variant uses estimated scores from finite rollout groups and removes the negative branch, so the comparison reflects both changes. Exponentiation can concentrate weight on a few rollouts when scores are noisy, whereas the centered, bounded update retains the first-order direction of the population tilt while limiting the influence of score outliers.

The initialization study tests whether the result relies on VideoAlign-specific reward weights. Replacing VideoAlign initialization with Qwen2-VL-Instruct gives 86.85 versus 87.55 Physics VBench, while General VBench rises from 60.83 to 61.25 and VideoAlign-MQ from 0.605 to 0.611. The gains therefore do not require VideoAlign initialization, although the lower VideoPhy-2-PC score (4.12 versus 4.20) shows that initialization still affects the balance across metrics.

Table 12: Ablations on the training objective and discriminator initialization.

Study	Variant	Physics VBench↑	General VBench↑	VideoAlign-MQ↑	VideoPhy-2-PC↑
Objective	Original bounded two-branch AFD	85.38	60.99	–	4.18
Objective	Exponential tilt	81.80	60.66	–	4.11
Initialization	VideoAlign	87.55	60.83	0.605	4.20
Initialization	Qwen2-VL-Instruct	86.85	61.25	0.611	4.12



Figure 8: Qualitative examples of AFD on a Disney-style animation dataset for continual learning. Detailed prompts are provided in Table 9.

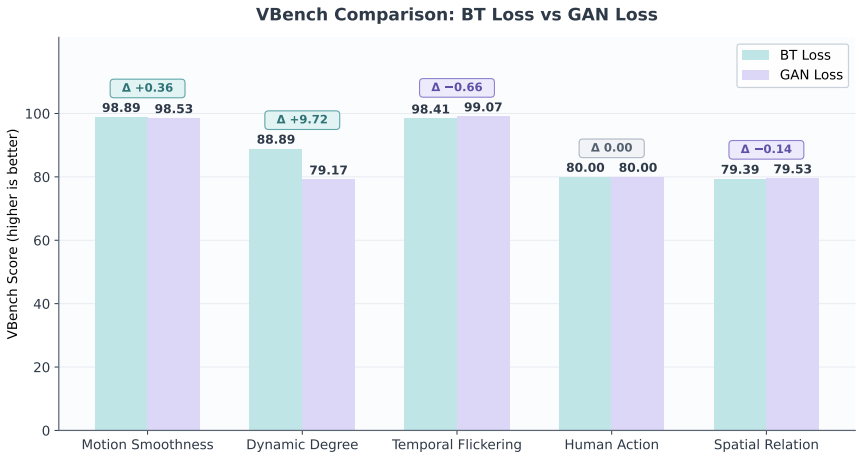

Figure 9: VBench discriminator-loss ablation. BT improves dynamic degree relative to the GAN-style binary loss, while the other shown dimensions remain close.

Pairwise versus binary discrimination. Figure 9 holds the student-side update fixed and changes the discriminator loss. BT raises Dynamic Degree from 79.17 to 88.89, while each of the other four shown VBench dimensions changes by less than one point. The benefit is therefore concentrated in generated motion rather than a uniform increase across metrics, supporting pairwise feedback for motion-sensitive adaptation.

G Limitations

Our experiments cover two causal AR students, one bidirectional student, and three sampling-only teachers. Most evaluations use physics-oriented prompts, with a smaller stylized set. Longer videos, broader prompt distributions, and additional teacher and positive-data sources remain to be tested. AFD's performance can depend on the discriminator update rate and reward calibration. The current ablations do not identify a generally optimal setting. The discriminator assigns feedback to complete

rollouts rather than individual frames or blocks. Teacher-generated positives provide a prompt-matched transfer source, but the present results do not establish that they are preferable to curated real or synthetic videos.