SHIFT: Motion Alignment in Video Diffusion Models with Adversarial Hybrid Fine-Tuning

Xi Ye[1,2], Wenjia Yang[1], Yangyang Xu[1,2], Xiaoyang Liu[1,3], Duo Su[1], Mengfei Xia[2], and Jun Zhu[1(✉)]

[1] Dept. of Computer Science and Technology, Tsinghua University, Beijing, China
dcszj@mail.tsinghua.edu.cn
[2] Ant Group, Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China

Abstract. Image-conditioned video diffusion models achieve impressive visual realism but often suffer from weakened motion fidelity, e.g., reduced motion dynamics or degraded long-term temporal coherence, especially after fine-tuning. We study motion alignment in video diffusion models post-training. To address this, we introduce pixel-motion rewards based on pixel flux dynamics, capturing both instantaneous and long-term motion consistency. We further propose Smooth Hybrid Fine-tuning (SHIFT), a scalable reward-driven framework that unifies supervised fine-tuning and advantage-weighted fine-tuning. Benefiting from novel adversarial advantages, SHIFT improves convergence speed and mitigates reward hacking. Experiments show that our approach efficiently resolves dynamic-degree collapse in modern video diffusion models supervised fine-tuning. Project page: https://xiye20.github.io/projects/SHIFT/.

Keywords: Video diffusion · Reward Fine-tuning · Motion alignment

1 Introduction

Recent advances in diffusion-based generative modeling have enabled high-quality video synthesis across image-to-video (I2V) and text-image-to-video (TI2V) [3, 31, 37]. By extending powerful image diffusion backbones with temporal modeling, modern video diffusion models produce visually compelling results with strong semantic fidelity. Yet high visual quality does not guarantee temporally faithful motion, and generated videos often lack rich dynamics.

Motion is fundamental to video realism. Beyond frame-level visual quality, realistic video generation requires coherent object trajectories, sustained motion magnitude, and physically plausible temporal evolution. However, many video diffusion models exhibit motion attenuation, where object movement becomes overly smooth, damped, or static. This issue is particularly pronounced after supervised fine-tuning (SFT) of image-conditioned models, I2V or TI2V models

X. Ye and W. Yang—Equal Contribution.

often suffer from a dynamic-degree degradation phenomenon: motion strength and long-term dynamics decrease even as visual sharpness and semantic alignment improve [30, 39, 41]. Recent benchmarking further suggests that visual realism does not imply physical understanding, and that current generative video models frequently violate basic physical principles [21].

Reinforcement learning (RL) is a typical post-training method to improve desired quality of video diffusion models. However, aligning video diffusion models with motion-centric objective presents several challenges. First, motion is a high-dimensional, low-level, and continuous signal over spatiotemporal space, making it difficult to supervise using semantic labels. Human feedback is poorly suited for evaluating subtle physical dynamics at scale, while existing vision-language models lack the precision required for pixel-level motion assessment. Second, motion quality depends on both short-term coherence and long-term temporal structure, requiring reward signals that operate across multiple time scales.

In this work, we argue that effective motion alignment should be grounded in low-level motion representations rather than high-level semantics. We propose to model pixel trajectories via a transport residual derived from the optical-flow equation, providing a low-level, motion-aware description of temporal dynamics. Based on this representation, we design two complementary motion rewards: an Instantaneous Motion reward (IMR) that enforces local temporal coherence between adjacent frames, and a Long-term Motion reward (LMR) that captures sustained dynamics and global motion structure over extended horizons. These rewards are dense, scalable, and directly tied to pixel-level motion, enabling fine-grained supervision without human annotations or semantic models. In this way, our pixel-motion reward model addresses the failure mode of motion degradation by explicitly incentivizing both motion magnitude and temporal coherence.

Beyond the reward design, applying RL to video diffusion models poses further challenges: MDP-based methods [2] are computationally expensive, tightly coupled to specific samplers, and difficult to scale; gradient-based alternatives [7, 26] require backpropagating through the full reverse process and cannot use black-box rewards. Moreover, optimizing imperfect rewards risks reward hacking, which is amplified by the long temporal horizons of video generation.

To overcome the aforementioned challenges of RL in video diffusion models, we introduce Smooth Hybrid Fine-Tuning (SHIFT) with adversarial advantage. SHIFT dynamically fuses SFT and reward-based fine-tuning into a stable unified optimization framework. Besides, SHIFT alternates between fine-tuning the diffusion model and updating the motion reward model. By explicitly incorporating adversarial training, SHIFT accelerates convergence, and mitigates reward hacking. Furthermore, unlike reverse-process reinforcement learning methods, SHIFT operates entirely in the forward diffusion process, does not require storing denoising trajectories, and remains agnostic to the choice of diffusion sampler.

Our contributions are threefold:

– We propose pixel-motion reward models that capture both instantaneous and long-term motion fidelity, enabling scalable motion supervision without human preference labels.

– We introduce a hybrid fine-tuning framework that unifies an offline SFT anchor with advantage-weighted updates on model rollouts in the forward process, remaining efficient and sampler-agnostic.
– Comprehensive evaluations and ablations demonstrating that SHIFT improves motion dynamics while preserving appearance quality.

2 Related works

Standard I2V/TI2V diffusion denoising objective provides no explicit incentive to preserve motion magnitude, causing supervised fine-tuning to bias models toward visually stable but physically inert solutions. Recent efforts to incorporate physics information into video generation [17] vary in both representation (low-level implicit vs. high-level symbolic) and integration strategy (conditional training vs. external feedback). Our motion-aware reward model falls in the low-level implicit category and serves as an external feedback signal for the SHIFT fine-tuning framework.

Motion quality in video diffusion models. Image-to-video diffusion models frequently suffer dynamic-degree degradation during supervised fine-tuning, where motion strength and long-term temporal coherence deteriorate even as visual quality improves [30, 39, 41]. Several directions have been explored to address this. Explicit geometric supervision methods such as Track4Gen [14] employ point-tracking losses for trajectory-level guidance but rely on high-level semantic alignment, failing to capture fine-grained physical motion laws [5, 11, 38]. Inference-time corrections [30] amplify motion via training-free model merging but do not rectify the model's underlying bias. Preference-based approaches like DenseDPO [27, 36] introduce segment-level labels, yet their reliance on human perception introduces subjectivity and favors frame visual quality over motion consistency. In contrast, our method provides precise dynamic feedback through motion-aware reward models that require no human intervention.

Reinforcement learning in diffusion models. RL-based post-training alignment of diffusion models encompasses MDP-based policy gradient methods [2, 8, 18], reward-weighted regression (RWR) [16, 22], and gradient-based techniques [7, 26] (see supplementary material for a detailed survey). Scaling these to video diffusion, however, remains challenging: MDP-based methods incur prohibitive memory costs from storing full denoising trajectories, both MDP and RWR approaches are tightly coupled to specific samplers, e.g., SDE sampler, and long temporal horizons exacerbate credit assignment [7, 35]. These difficulties are compounded by reward hacking, where generators exploit imperfections in reward models rather than learning intended behaviors, which is especially severe in high-dimensional video generation [1, 34]. Existing mitigations, including KL regularization [13], GRPO [18, 29], and adversarial reward updates [20, 40], address this partially. In contrast, SHIFT operates entirely in the forward process with online adversarial advantages, requiring neither human feedback nor specific diffusion samplers.

3 Preliminaries: Diffusion RL and Limitations.

The goal of reinforcement learning in diffusion is to fine-tune a pre-trained model p_θ to maximize a reward function $r(x_0)$, which reflects human preference or specific downstream objectives, and x_0 denotes a clean data example. Denoising Diffusion Policy Optimization (DDPO) [2] formulates the reverse diffusion sampling process as a multi-step Markov Decision Process (MDP). Thus, DDPO requires an SDE-solver for sampling. The sampling process forms a trajectory $\tau = \{x_T, \ldots, x_0\}$, where each denoising step $p_\theta(x_{t-1}|x_t)$ represents the policy action, and a reward is observed only at the final state x_0.

To ensure optimization stability, trust-region methods (e.g., PPO) typically impose a *Reverse KL* constraint between the updated policy p_θ and a reference policy p_{ref} (usually the pre-trained model):

$$\max_\theta \mathbb{E}_{x_{0:T}\sim p_\theta}[r(x_0)] \quad \text{s.t.} \quad D_{\text{KL}}(p_\theta||p_{\text{ref}}) \leq \gamma. \tag{1}$$

The closed-form optimal solution to this constrained optimization problem is known to take the form of an energy-based distribution [22, 23]:

$$p^*(x) \propto p_{\text{ref}}(x) \exp(\frac{r(x)}{\beta}), \tag{2}$$

where β is a temperature parameter balancing reward maximization and the KL constraint. Standard approaches optimize this objective via policy gradients. The gradient of the regularized objective is derived as:

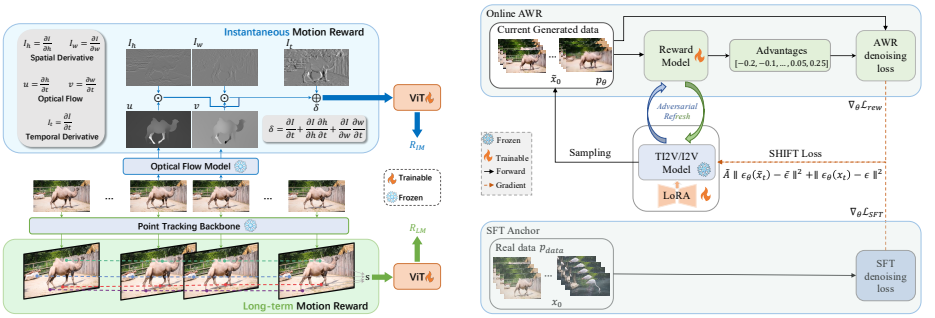
$$\nabla_\theta \mathcal{L}_{\text{RL}} = -\mathbb{E}_{x_{0:T}}\left[\nabla_\theta \log p_\theta(x_{0:T}) \cdot r(x_0)\right] + \nabla_\theta D_{\text{KL}}(p_\theta||p_{\text{ref}}) \tag{3}$$

$$= -\mathbb{E}_{x_{0:T}}[\sum_{t=1}^{T} \nabla_\theta \log p_\theta(x_{t-1}|x_t) \cdot r(x_0)] + \nabla_\theta D_{\text{KL}}(p_\theta||p_{\text{ref}}). \tag{4}$$

This formulation highlights the dependence on full trajectory optimization. For the KL-div term in Eq.(3), we can derive a closed-form solution, please refer to [18] for more detail. Empirically, prior works utilize the relative group advantage $A(x_0)$ instead of raw reward $r(x_0)$ for a more stable training [18]. While effective for lightweight image models, applying PPO-style RL to large-scale video diffusion models presents significant challenges:

1. Computational Prohibitiveness of On-Policy Sampling. PPO-style updates require frequent on-policy sampling to keep the importance sampling ratio close to 1. For high-resolution video models, generating fresh rollouts is computationally prohibitive. Stale data leads to divergent importance weights and high-variance updates, creating a dilemma: we cannot afford frequent online sampling, yet cannot train stably without it.

2. Structural Mismatch between MDP and Diffusion. Formulating diffusion fine-tuning as a standard MDP imposes significant solver restrictions and discretization inconsistencies. While diffusion models are trained to approximate continuous marginal vector fields (via score or flow matching) to support



(a) Reward models (b) Smooth Hybrid Fine-Tuning framework

Fig. 1: Overview of the proposed approach. (a) Pixel-motion reward models: instantaneous reward based on optical-flow residual (top) and long-term reward based on trajectory dynamics (bottom). (b) The SHIFT fine-tuning framework.

flexible inference, MDP-based RL methods lock the training process to specific discrete reverse chains. This introduces a "forward-reverse inconsistency" where the fine-tuning objective diverges from the pre-training paradigm [4]. Furthermore, optimizing discrete trajectories complicates integration with inference-time techniques like Classifier-Free Guidance (CFG) [4, 42].

4 Methodology

In this work, we fine-tune pre-trained video diffusion models to enhance *motion fidelity*, implicitly defined by pixel-level motion consistency. We posit that real video distributions inherently exhibit correct motion dynamics. To align the generated distribution with this prior, we introduce a motion-aware reward mechanism based on optical-flow features, operating across both instantaneous and long-term temporal scales, see Figure 1(a).

We then propose **SHIFT** (Smooth Hybrid Fine-tuning, Figure 1(b)), a framework that fuses Advantage-Weighted Regression (AWR) on generated samples with an SFT anchor on real data to ensure stability. To mitigate reward hacking, we adopt an adversarial paradigm, alternating between diffusion model updates and reward model refinement.

4.1 Pixel-Motion Reward Models

Instantaneous Motion Reward (IMR). To capture local temporal coherence, we model video as a continuous scalar field of pixel intensity $I(\mathbf{x}, t)$: $\Omega \times \mathbb{R}^+ \to \mathbb{R}$, defined over the spatial image domain $\Omega \subset \mathbb{R}^2$ with coordinates $\mathbf{x} = (h, w) \in \Omega$ and time t. Inspired by fluid dynamics, the evolution of intensity is related to a continuity equation, which decomposes temporal changes into

distinct motion components:

$$S = \frac{\partial I}{\partial t} + \nabla \cdot (I\mathbf{u}) = \frac{\partial I}{\partial t} + \underbrace{\mathbf{u} \cdot \nabla I}_{\text{Advection}} + \underbrace{I(\nabla \cdot \mathbf{u})}_{\text{Divergence}}, \tag{5}$$

where $\mathbf{u} = (u, v) = (\frac{\partial h}{\partial t}, \frac{\partial w}{\partial t})$ represents the pixel velocity field. The advection term, $\mathbf{u} \cdot \nabla I$, captures the transport of pixel values along motion trajectories (e.g., object translation). The divergence term, $I(\nabla \cdot \mathbf{u})$, accounts for intensity variations induced by flow compression or expansion.

To adapt this motion prior to video generation, we must account for the distinct properties of visual data. Unlike compressible fluids where expansion necessitates density reduction to conserve mass, pixel intensity represents surface radiance, which remains stable under field-of-view changes. For example, during a camera zoom-in ($\nabla \cdot \mathbf{u} > 0$), object brightness does not decrease. Enforcing the full continuity equation would incorrectly penalize such natural geometric scaling unless the video artificially darkened. We therefore refine the motion constraint by omitting the divergence term, approximating the transport residual δ via the advective component alone:

$$\delta \approx \frac{\partial I}{\partial t} + \mathbf{u} \cdot \nabla I = \frac{\partial I}{\partial t} + \frac{\partial h}{\partial t}\frac{\partial I}{\partial h} + \frac{\partial w}{\partial t}\frac{\partial I}{\partial w}. \tag{6}$$

While $\delta = 0$ corresponds to the classical brightness constancy assumption [10], we empirically observe that directly minimizing $\|\delta\|^2$ yields pathological, such as blurry and static motion. Real-world videos exhibit non-zero residuals due to complex lighting and sensor noise. Nevertheless, δ serves as a critical motion-discriminative feature: its statistical distribution differs significantly between realistic motion and generated artifacts (e.g., high-frequency flickering).

Instead of forcing $\delta \to 0$, we treat it as a dense feature map. As illustrated in Figure 1(a), we compute the components of Eq. (6) using differentiable operators. Spatial gradients $\frac{\partial I}{\partial h}$ and $\frac{\partial I}{\partial w}$ are extracted via Sobel operators, while the temporal derivative $\frac{\partial I}{\partial t}$ is approximated by finite differences between adjacent frames. The velocity field components, $u = \frac{\partial h}{\partial t}$ and $v = \frac{\partial w}{\partial t}$, are estimated using a pre-trained SEA-RAFT model [33]. To account for estimation errors in the flow field, we concatenate the computed residual map δ with the flow confidence map provided by SEA-RAFT. This fused representation is processed by a Vision Transformer (ViT) trained to discriminate between the natural motion residuals of real videos and the artifacts of generated samples, thereby providing a robust, motion-discriminative signal for local temporal coherence.

Long-term Motion Reward (LMR). Beyond local coherence, motion fidelity requires consistent global dynamics over extended horizons. As illustrated in Figure 1(a), we evaluate long-term motion by tracking point trajectories $\tau_{\mathcal{P}} \in \mathbb{R}^{T \times M \times 2}$ initialized from a set of M query points $\mathcal{P}$ sampled randomly from the first frame of the input video $\mathbf{X} \in \mathbb{R}^{T \times 3 \times H \times W}$. Unlike fixed grid sampling, this stochastic query initialization acts as an implicit ensemble over motion constraints, enhancing robustness against reward hacking.

We treat the evolution of pixel trajectories as a temporal stochastic process. To incorporate this motion-based inductive bias, we explicitly model the drift component via the instantaneous velocity of tracked points, $\tau_d[t] = \tau[t] - \tau[t-1]$. Trajectories are estimated using CoTracker3 [15], which additionally provides visibility masks τ_v and confidence scores τ_c. These components are concatenated to form a motion state descriptor $\mathbf{s} = [\tau_d, \tau_v, \tau_c]$. To further capture global structural dependencies, we augment this descriptor with dense correlation maps representing spatial affinities between query points and their temporal neighbors.

Reward Training. Both the instantaneous and long-term reward models are parameterized as Vision Transformers (ViTs) and trained as binary discriminators to distinguish between real videos drawn from the dataset p_{data} and synthetic samples generated by the diffusion model. Formally, we optimize the discriminator parameters ω by minimizing the binary cross-entropy loss:

$$\mathcal{L}_{\text{rew}}(\omega) = -\mathbb{E}_{x\sim p_{\text{data}}}[\log D_\omega(x)] - \mathbb{E}_{\tilde{x}\sim p_\theta}[\log(1 - D_\omega(\tilde{x}))]. \tag{7}$$

The reward $r(\mathbf{x})$ is defined as the raw logit output of the discriminator, $r(\mathbf{x}) = \text{logit}(D_\omega(\mathbf{x}))$, clipped to a stable range $[r_{\min}, r_{\max}]$. We explicitly opt for raw logits over sigmoid-transformed probabilities to avoid gradient saturation, ensuring stronger learning signals particularly when the discriminator is confident (see ablation study).

4.2 SHIFT: Smooth Hybrid Fine-tuning.

To address aforementioned challenges in section 3, we present **SHIFT** (**S**mooth **H**ybr**I**d **F**ine-**T**uning), illustrated in Figure 1(b), a data-regularized reinforcement learning framework designed for efficient video diffusion alignment. Instead of constraining the model to a drifting reference policy via Reverse KL, we anchor the model to the stationary ground-truth data distribution p_{data} using *Forward KL*. This is particularly valid for motion alignment, where we assume ground-truth video data possesses perfect motion fidelity.

Deriving the Forward-KL Target. Replacing the reference policy in Eq. (2) with stationary data distribution, we define optimal target $p^*(x)$ as:

$$p^*(x) \propto p_{\text{data}}(x) \exp(\frac{r(x)}{\beta}). \tag{8}$$

While minimizing the Reverse KL between this target and p_θ is intractable due to the unnormalized density of p^*, minimizing the **Forward KL** divergence is straightforward:

$$\mathcal{J}(\theta) = D_{\text{KL}}(p^*(x)||p_\theta(x)) \equiv \arg\max_\theta \mathbb{E}_{x\sim p^*}[\log p_\theta(x)]. \tag{9}$$

Optimization via Diffusion Loss Proxy. Evaluating $\mathbb{E}_{x\sim p^*}[\dots]$ is non-trivial because we cannot directly sample from p^*. However, since p^* is a product of the data distribution and the reward exponent, its probability mass concentrates in two regions: (1) the support of p_{data}, and (2) regions of high reward

Algorithm 1 SHIFT: Smooth Hybrid Fine-tuning

Input: pre-trained denoiser ϵ_{base}, reward model r_ω, dataset p_{data}, iterations N, rollout size B, inner steps K, temperature β.
Initialize: $\theta \leftarrow \theta_{\text{base}}$.

for $n = 1$ to N **do**
  Sample real data $\{x_0^b\}_{b=1}^B \sim p_{\text{data}}$ and fake samples $\{\tilde{x}_0^b\}_{b=1}^B \sim p_\theta$.
  Compute scaled rewards $\tilde{r}^b = r_\omega(\tilde{x}_0^b)/\beta$, $r^b = r_\omega(x_0^b)/\beta$, and mean $\bar{r} = \frac{1}{B}\sum_b \tilde{r}^b$.
  Compute advantages $\tilde{A}^b = \tilde{r}^b - \bar{r}$, $A^b = r^b - \bar{r}$.
  for $k = 1$ to K **do**
    Update reward model by $\mathcal{L}_{\text{rew}} = -\mathbb{E}_b[\log D_\omega(x_0^b) + \log(1 - D_\omega(\tilde{x}_0^b))]$.
    Sample $t \sim \mathcal{U}(0, T)$, $x_t \sim q(x_t|x_0)$, $\tilde{x}_t \sim q(x_t|\tilde{x}_0)$.
    Update generator with

$$\mathcal{L}_{\text{SHIFT}} = \mathbb{E}_{b,t}\Big[\tilde{A}^b \|\epsilon_\theta(\tilde{x}_t, t) - \tilde{\epsilon}\|^2 + \|\epsilon_\theta(x_t, t) - \epsilon\|^2\Big].$$

  end for
end for

$r(x)$. We therefore approximate the gradient $\mathbb{E}_{x\sim p^*}[\nabla_\theta \log p_\theta(x)]$ with a mixture estimator over two accessible sources. An *offline anchor* ($x \sim p_{\text{data}}$) samples directly from the dataset to preserve the generation quality and motion coherence of the base distribution, while an *online exploration* term ($\tilde{x} \sim p_\theta$) uses the model's own rollouts, weighted by their relative advantage $A(\tilde{x})$, to cover high-reward regions favored by $\exp(r(\tilde{x})/\beta)$. This yields:

$$\mathcal{J}(\theta) \approx \underbrace{\mathbb{E}_{\tilde{x}\sim p_\theta}[\exp(\frac{A(\tilde{x})}{\beta}) \log p_\theta(\tilde{x})]}_{\text{Online Exploration}} + \underbrace{\mathbb{E}_{x\sim p_{\text{data}}}[\log p_\theta(x)]}_{\text{Offline Anchor}}. \tag{10}$$

Approximating Likelihood with Diffusion ELBO. We avoid the high memory cost of optimizing reverse trajectories by approximating the log-likelihood gradient $\nabla_\theta \log p_\theta(x)$ with the gradient of the Variational Lower Bound (ELBO). It is well-established that maximizing the log-likelihood is approximately equivalent to minimizing the weighted denoising error (please see the supplementary material for the detailed derivation):

$$\nabla_\theta \log p_\theta(x) \approx -\nabla_\theta \mathcal{L}_{\text{diff}}(x) = -\nabla_\theta \mathbb{E}_{t,\epsilon}[\|\epsilon_\theta(x_t, t) - \epsilon\|^2]. \tag{11}$$

Substituting this proxy into our hybrid objective yields the final loss function:

$$\mathcal{L}_{\text{SHIFT}} = \mathbb{E}_{b,t}[\underbrace{\tilde{A}^b \|\epsilon_\theta(\tilde{x}_t, t) - \tilde{\epsilon}\|^2}_{\text{Online RL Update}} + \underbrace{\|\epsilon_\theta(x_t, t) - \epsilon\|^2}_{\text{Offline Data Anchor}}]. \tag{12}$$

Here, the first term reinforces self-generated trajectories ($\tilde{x}$) that achieve high rewards ($\tilde{A}^b > 0$), while the second term acts as a Supervised Fine-Tuning (SFT) regularizer preventing the policy from drifting out-of-distribution. This formulation provides a smooth integration of RL and SFT: when the advantage

is zero, the algorithm reduces to standard fine-tuning; when rewards are high, it aggressively shifts the distribution toward desirable metrics. Detailed derivations of Eq. (12) are provided in the supplementary material.

4.3 Adversarial Advantages

Advantages. Similar to the GRPO [29], at each iteration, we sample a mini-batch of real videos $\{x_0^b\}_{b=1}^B \sim p_{\text{data}}$ and generate $\{\tilde{x}_0^b\}_{b=1}^B \sim p_\theta$ using a fixed discrete sampler. Then we compute temperature-scaled rewards $\tilde{r}^b = r_\omega(\tilde{x}_0^b)/\beta$ and $r^b = r_\omega(x_0^b)/\beta$, and center using the fake mean $\bar{r} = \frac{1}{B}\sum_b \tilde{r}^b$: $\tilde{A}^b = \tilde{r}^b - \bar{r}$ and $A^b = r^b - \bar{r}$. We use only recentering (no standard-deviation normalization) to avoid difficulty bias induced by small reward variance [19].

Adversarial Online discriminator update. In unverifiable generation tasks, the generator distribution p_θ inevitably drifts from the initial training support, rendering static rewards unreliable and susceptible to hacking. To mitigate this, we employ an iterative adversarial update strategy. Periodically, we refresh the discriminator by keep optimizing Eq. (7), using the most recent on-policy samples $\tilde{x}_0^b$ as negative examples against the ground-truth videos. In this framework, the reward approximates the log-density ratio $\log \frac{p_{\text{data}}(\mathbf{x})}{p_\theta(\mathbf{x})}$ in feature space, ensuring the signal remains accurate and robust to distribution shift throughout training. The complete SHIFT procedure is outlined in Algorithm 1.

Remark. While SHIFT's loss structure (Eq. (12)) bears a superficial resemblance to the reward-weighted regression plus SFT objective of Lee et al. [16], the two methods differ in several fundamental respects. First, SHIFT derives its objective from a principled Forward KL formulation (Eq. (9)), whereas Lee et al. adopt a heuristic combination of RWR and SFT without such theoretical grounding. Second, SHIFT replaces raw rewards with group-relative advantages and employs adversarial reward-model co-training to prevent reward hacking—mechanisms absent in Lee et al. Third, our reward models are fully automatic, motion-aware discriminators trained without any human annotation; this is essential for video motion alignment, where pixel-level temporal dynamics are practically infeasible for humans to label at scale. Finally, although SHIFT alternates between generator and discriminator updates, it differs fundamentally from standard GAN training: the discriminator provides only a scalar reward signal used to compute advantages, and its gradient is never backpropagated into the generator. As shown by Pfau and Vinyals [24], GANs can be viewed as actor-critic methods where coupled gradient-based min-max optimization is the root cause of training instability; SHIFT sidesteps this by fully decoupling the two optimizations, retaining the discriminator's adaptive distribution-awareness without inheriting adversarial gradient dynamics.

5 Experiments

Implementation. We validate SHIFT on two video diffusion models: SVD [3] (∼1.2B parameters), fine-tuned on DAVIS2017 [25] (∼4K training clips after

Table 1: Quantitative results on VBench-I2V for the SVD model. FVD are tested on the DAVIS2017 validation set. Higher (↑) is better; lower (↓) is better. Best is **bold**; second-best is underlined.

	VBench Appearance (↑)	VBench Motion (↑)	VBench Overall (↑)	Motion Score (↑)	FVD (↓)	Training Time (↓)
Base	83.68	85.89	84.41	4.39	395.87	-
SFT	**85.07**	76.09	82.08	2.69	322.81	**1x**
Track4Gen [14]	85.01	76.44	82.15	2.84	**316.91**	1.07x
DenseDPO [35]	83.73	86.03	84.50	4.34	399.73	2.40x
FlowGRPO [18]	83.61	86.67	84.63	**4.44**	396.56	19.35x
SHIFT (Ours)	83.69	**86.70**	**84.69**	4.40	396.45	2.46x

temporal segmentation), following the standard SVD fine-tuning protocol established by Track4Gen [14], and the Wan2.2 TI2V model (∼5B parameters) [31], fine-tuned on WISA-80K [32]. For SVD, LoRA modules (rank 32) are applied only to temporal attention layers; for Wan2.2, LoRA (rank 32) is applied to all attention layers. Notably, unlike MDP-based baselines (e.g., FlowGRPO) that require stochastic samplers, SHIFT is sampler-agnostic. Full training and sampling configurations are provided in the supplementary material.

Metrics. We adopt VBench-I2V [12], grouping its sub-metrics into appearance-related (subject/background consistency, aesthetic/imaging quality, I2V subject/background) and motion-related (temporal flickering, motion smoothness, dynamic degree). We report category averages and the overall score on standard VBench-I2V, along with motion score [41] and Fréchet Video Distance (FVD).

Reward model pre-training. Prior to RL fine-tuning, we pre-train the motion reward models as binary discriminators on paired real and generated videos. For each real video, we extract its first frame (optionally with the text prompt) and sample four synthetic videos from the base model as negatives. Details on data balancing and convergence monitoring are provided in the supplementary material.

5.1 SVD fine-tuning

For a fair comparison, all post-training methods are applied directly to the base model on the same hardware (64 GPUs), with per-epoch wall-clock time normalized to that of SFT (1×); we set the buffer reuse factor K=10 for both SHIFT and FlowGRPO. Table 1 summarizes the results.

SFT and Track4Gen [14] achieve the lowest FVD scores (322.81 and 316.91), but at a severe cost to motion quality: VBench Motion drops by 9.80 and 9.45 points, and Motion Score falls from 4.39 to 2.69 and 2.84 respectively, confirming the dynamic-degree degradation discussed in Section 1. Per-dimension analysis shows that this is driven mostly by the collapse of the dynamic degree metric (from 0.67 to 0.33 for SFT), indicating near-static video generation. As noted by Ge et al. [9], FVD is highly insensitive to temporal quality and biased toward per-frame appearance, rendering these seemingly strong FVD numbers misleading.

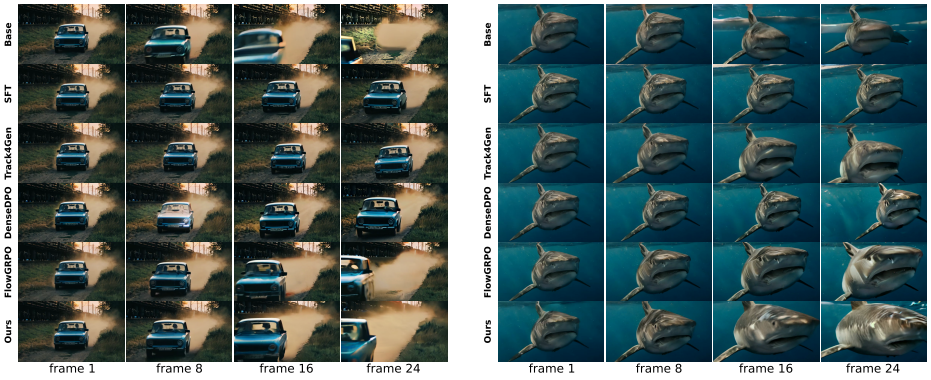


Fig. 2: Qualitative comparison of VBench-I2V Standard test examples generated by the base SVD model and different fine-tuned SVD variants.

DenseDPO [35] preserves motion quality (VBench Motion 86.03, Motion Score 4.34) but slightly worsens FVD (399.73 vs. 395.87) and yields only marginal overall gains (+0.09). FlowGRPO [18] further improves motion, achieving the highest Motion Score (4.44) and strong gains in VBench Motion (86.67, +0.78) and Overall (84.63, +0.22); however, it slightly degrades appearance (83.61 vs. 83.68) and is considerably more expensive (19.35× vs. 2.46×). In contrast, SHIFT achieves the best VBench Motion (86.70, +0.03 over FlowGRPO) and Overall (84.69, +0.06 over FlowGRPO) while maintaining appearance on par with the base model (83.69 vs. 83.68) and comparable FVD (396.45). SHIFT is the only method that simultaneously improves motion fidelity, overall quality, and motion magnitude without degrading any individual metric, at a fraction of FlowGRPO's computational budget.

Regarding training efficiency, Track4Gen adds a lightweight correspondence loss atop SFT (1.07×). DenseDPO (2.40×) and SHIFT (2.46×) incur comparable costs from winner–loser pair processing and adversarial reward-model updates, respectively; both operate entirely in the forward diffusion process without optimizing through the reverse denoising trajectory. FlowGRPO is substantially more expensive (19.35×, ≈7.9× that of SHIFT) because it formulates denoising as a Markov decision process and recomputes per-timestep log-probabilities across all 25 sampling steps, requiring much more UNet forward passes per step.

Please see Figure 2 for qualitative examples showing that SHIFT improves motion dynamics while preserving strong visual quality. More comparisons are provided in the supplementary material.

5.2 Wan2.2-TI2V fine-tuning

We further evaluate SHIFT on the 5B Wan2.2-TI2V model fine-tuned on the deformation subset of WISA-80K, which contains relatively rare real-world physical phenomena. Table 2 shows a trend consistent with the SVD experiments: SFT

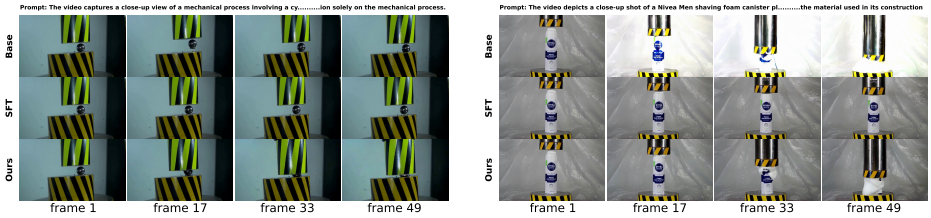


Fig. 3: Qualitative comparison on the WISA-80K validation set for the base Wan2.2-TI2V model and fine-tuned variants. SHIFT generates more realistic motion.

Table 2: Quantitative results on VBench-I2V Standard for the Wan2.2 TI2V model. Higher (↑) is better; lower (↓) is better. Best is **bold**; second-best is underlined.

	VBench Appearance (↑)	VBench Motion (↑)	VBench Overall (↑)	Motion Score (↑)	FVD (↓)
Base	86.29	86.25	86.27	**2.96**	284.68
SFT	**87.14**	83.34	85.87	1.94	**151.44**
SHIFT (ours)	86.38	**86.45**	**86.40**	2.93	282.24

improves appearance and FVD but substantially degrades motion (Motion Score 1.94 vs. 2.96). In contrast, SHIFT improves the overall VBench score (86.40 vs. 86.27) with modest gains in both appearance and motion while keeping FVD comparable to the base model. Figure 3 provides qualitative examples, and additional results are included in the supplementary material.

5.3 Ablation study

We analyze the contribution of each SHIFT component by progressively adding regularization mechanisms to the Instantaneous Motion Reward baseline (A0 in Table 3): first Noise Alignment (A1), then Adversarial Reward Model updates (A2). These two mechanisms serve primarily to stabilize training and prevent reward hacking rather than to boost motion quality directly. Since the Long-term Motion Reward provides an orthogonal motion supervision signal, we evaluate it separately by adding LMR to the A1 baseline (A3). The full model (A4) combines all four components. Figure 4 summarizes the training dynamics.

IMR baseline (A0). Training with IMR alone yields the highest VBench Motion among all variants (91.01) but at a steep cost: Appearance falls to 81.62 and FVD rises to 502.60, indicating reward hacking. As shown in Figure 4, this degradation worsens over continued training. We attribute this to a *noise-level mismatch* between the SFT anchor and the advantage-weighted regression (AWR) term: higher noise levels predominantly govern coarse structure and motion, while lower levels control fine visual details [6]. When the two loss terms sample noise levels independently, AWR aggressively optimizes motion at high noise scales while SFT regularization is diluted across random scales, leaving appearance unprotected.

Table 3: Ablation of SHIFT components evaluated at the same epoch 7. Higher (↑) is better; lower (↓) is better. Best is **bold**; second-best is underlined.

	IMR	Noise Align	Adv. RM	LMR	VBench Appearance (↑)	VBench Motion (↑)	VBench Overall (↑)	Motion Score (↑)	FVD (↓)
A0	✓				81.62	**91.01**	84.75	4.29	502.60
A1	✓	✓			82.53	87.18	84.08	3.75	472.20
A2	✓	✓	✓		83.06	87.28	84.47	3.75	451.20
A3	✓	✓		✓	82.48	89.31	**84.76**	4.14	460.00
A4	✓	✓	✓	✓	**83.63**	86.98	84.75	**4.35**	**401.76**

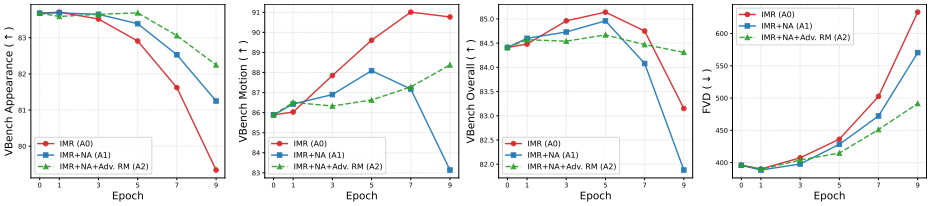


Fig. 4: Ablation of SHIFT components over training epochs. All models start from the same pre-trained SVD checkpoint. Adding Noise Alignment (NA) strengthens the SFT anchor regularization, while the Adversarial RM further stabilizes training to avoid reward hacking.

+ **Noise Alignment (A1).** To address this mismatch, Noise Alignment (NA) couples the noise schedules of SFT and AWR by sharing the same sampled noise level within each training step. As shown in Table 3, NA improves Appearance to 82.53 (+0.91 over A0) and reduces FVD to 472.20 (vs. 502.60). Figure 4(a, d) further confirms that NA produces more stable training dynamics with substantially less reward hacking, confirming that aligning noise strengthens the SFT anchor at scales where the AWR objective exerts strong gradient.

+ **Adversarial Reward Model (A2).** While NA mitigates the noise-level mismatch, the static reward model becomes increasingly exploitable as the generator distribution shifts. Adding adversarial reward-model updates, which periodically optimizes the discriminator on the latest generated samples and real samples, further improves results: Appearance rises to 83.06, FVD decreases to 451.20, and VBench Overall reaches 84.47, outperforms A1. Figure 4 confirms that A2 exhibits the most stable training trajectory, demonstrating that adversarial co-training enables sustained improvement by mitigating reward hacking.

+ **Long-term Motion Reward (A3).** Adding LMR to the A1 baseline recovers the motion quality that NA alone sacrifices for stability. A3 improves VBench Motion by +2.13 and Motion Score by +0.39 over A1, while preserving comparable appearance (82.48 vs. 82.53). Notably, the dynamic degree—the VBench dimension most indicative of meaningful object motion—rises from 68.54 to 75.93, confirming that LMR's trajectory-level supervision prevents the model from satisfying the instantaneous reward through superficial artifacts (e.g., texture flickering) rather than genuine dynamics. A3 also achieves the best

VBench Overall (84.76), demonstrating that IMR and LMR provide complementary signals: IMR enforces local temporal coherence via optical-flow consistency, while LMR anchors global trajectory plausibility over the full horizon.

Full model (A4). Combining all four components yields the full SHIFT model, which largely prevents reward hacking and achieves the best overall balance between appearance and motion. Compared to A1, A4 recovers +1.10 in Appearance (83.63 vs. 82.53) and +0.60 in Motion Score (4.35 vs. 3.75), while reducing FVD by 70.44. Compared to A3, A4 further improves Appearance by +1.15 and reduces FVD by 58.24, confirming that adversarial reward training is essential for preventing distributional drift even when LMR is present. A4 attains the best Appearance, Motion Score, and FVD among all variants, trailing A3 in VBench Overall by only 0.01. These results reveal a clear division of roles: IMR and LMR provide the motion supervision signal at complementary temporal scales, while NA and Adv. RM serve as regularization mechanisms that prevent the optimization from degenerating into reward hacking. Neither NA nor Adv. RM improves motion quality per se; instead, they ensure that the motion gains from IMR and LMR translate into genuine quality improvements rather than distributional drift.

Logit vs. Sigmoid Reward. Our reward models are binary discriminators producing unbounded logits. We compare using the raw logit $r(\mathbf{x}) = f_\omega(\mathbf{x})$ versus the sigmoid probability $r(\mathbf{x}) = \sigma(f_\omega(\mathbf{x}))$ as the reward signal. The sigmoid saturates near 0 or 1 for large-magnitude logits, compressing the group-relative advantages $\tilde{A}^b = \tilde{r}^b - \bar{r}$ to near-zero variance and effectively stalling LoRA optimization. Raw logits preserve the full dynamic range of the discriminator's confidence, yielding well-separated advantages and meaningful gradient updates. Since raw logits are unbounded, we clip the computed advantages to a fixed range for stability, analogous to the clipping in PPO [28] and GRPO [29]. We adopt clipped raw-logit rewards throughout all experiments.

Additional ablations on temperature β and replay factor K are provided in the supplementary material. In summary, smaller β accelerates reward optimization but increases reward hacking, while moderate β offers a better stability–quality tradeoff. Reducing K from 10 to 5 causes minor metric changes but increases wall-clock cost, indicating that SHIFT is robust to moderate off-policy replay. We also ablate the reward model architecture (ViT vs. ResNet) and the LMR query point sampling strategy (uniform grid vs. random sampling) in the supplementary material.

6 Conclusion

We address dynamic-degree degradation in image-conditioned video diffusion models by introducing pixel-motion rewards derived from optical-flow and point-track features, capturing both instantaneous temporal coherence and long-term global motion dynamics. Built on these rewards, SHIFT unifies supervised anchoring with advantage-weighted updates in a sampler-agnostic framework, and incorporates adversarial reward-model co-training and noise-level alignment to

mitigate reward hacking. Experiments show that SHIFT improves motion fidelity without sacrificing appearance quality.

Acknowledgements. This work was supported in part by the NSFC under Grant 92570001; in part by the NSFC Joint Fund and Special Fund under Grants U25B6003 and 62550004. This work was supported by Ant Group Research Fund.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. In: The Twelfth International Conference on Learning Representations (2024)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Chen, H., Zheng, K., Zhang, Q., Cui, G., Cui, Y., Ye, H., Lin, T.Y., Liu, M.Y., Zhu, J., Wang, H.: NFT: Bridging supervised learning and reinforcement learning in math reasoning. In: The Fourteenth International Conference on Learning Representations (2026)
5. Chen, T.S., Lin, C.H., Tseng, H.Y., Lin, T.Y., Yang, M.H.: Motion-conditioned diffusion model for controllable video synthesis. arXiv preprint arXiv:2304.14404 (2023)
6. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11472–11481 (2022)
7. Domingo-Enrich, C., Drozdzal, M., Karrer, B., Chen, R.T.Q.: Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In: The Thirteenth International Conference on Learning Representations (2025)
8. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In: Advances in Neural Information Processing Systems. vol. 36, pp. 79858–79885 (2023)
9. Ge, S., Mahapatra, A., Parmar, G., Zhu, J.Y., Huang, J.B.: On the content bias in frechet video distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7277–7288 (2024)
10. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981)
11. Hu, Y., Chen, Z., Luo, C.: Lamd: Latent motion diffusion for image-conditional video generation. International Journal of Computer Vision **133**(7), 4384–4400 (2025)
12. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)

13. Jaques, N., Ghandeharioun, A., Shen, J.H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., Picard, R.: Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. arXiv preprint arXiv:1907.00456 (2019)
14. Jeong, H., Huang, C.H.P., Chul Ye, J., Mitra, N.J., Ceylan, D.: Track4gen: Teaching video diffusion models to track points improves video generation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7276–7287 (2025)
15. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6013–6022 (2025)
16. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
17. Lin, M., Wang, X., Wang, Y., Wang, S., Dai, F., Ding, P., Wang, C., Zuo, Z., Sang, N., Huang, S., et al.: Exploring the evolution of physics cognition in video generation: A survey. arXiv preprint arXiv:2503.21765 (2025)
18. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., ZHANG, D., Ouyang, W.: Flow-GRPO: Training flow matching models via online RL. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
19. Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W.S., Lin, M.: Understanding r1-zero-like training: A critical perspective. In: Second Conference on Language Modeling (2025)
20. Ma, Y., Klabjan, D., Utke, J.: Video to video generative adversarial network for few-shot learning based on policy gradient. IEEE Transactions on Neural Networks and Learning Systems pp. 1–15 (2025)
21. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? arXiv preprint arXiv:2501.09038 (2025)
22. Nair, A., Gupta, A., Dalal, M., Levine, S.: Awac: Accelerating online reinforcement learning with offline datasets. arXiv preprint arXiv:2006.09359 (2021)
23. Peters, J., Schaal, S.: Reinforcement learning by reward-weighted regression for operational space control. In: Proceedings of the 24th international conference on Machine learning. pp. 745–750 (2007)
24. Pfau, D., Vinyals, O.: Connecting generative adversarial networks and actor-critic methods. arXiv preprint arXiv:1610.01945 (2016)
25. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
26. Prabhudesai, M., Mendonca, R., Qin, Z., Fragkiadaki, K., Pathak, D.: Video diffusion alignment via reward gradients. arXiv preprint arXiv:2407.08737 (2024)
27. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems. vol. 36, pp. 53728–53741 (2023)
28. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
29. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
30. Tian, J., Qu, X., Lu, Z., Wei, W., Liu, S., Cheng, Y.: Extrapolating and decoupling image-to-video generation models: Motion modeling is easier than you think.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12512–12521 (June 2025)
31. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
32. Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al.: Wisa: World simulator assistant for physics-aware text-to-video generation. arXiv preprint arXiv:2503.08153 (2025)
33. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: European Conference on Computer Vision. pp. 36–54. Springer (2024)
34. Wu, J., Gao, Y., Ye, Z., Li, M., Li, L., Guo, H., Liu, J., Xue, Z., Hou, X., Liu, W., Zeng, Y., Huang, W.: Rewarddance: Reward scaling in visual generation. arXiv preprint arXiv:2509.08826 (2025)
35. Wu, Z., Kag, A., Skorokhodov, I., Menapace, W., Mirzaei, A., Gilitschenski, I., Tulyakov, S., Siarohin, A.: DenseDPO: Fine-grained temporal preference optimization for video diffusion models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
36. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., Teng, J., Yang, Z., Zheng, W., Liu, X., Zhang, D., Ding, M., Zhang, X., Gu, X., Huang, S., Huang, M., Tang, J., Dong, Y.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2026)
37. Ye, X., Bilodeau, G.A.: Vptr: Efficient transformers for video prediction. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 3492–3499. IEEE (2022)
38. Ye, X., Bilodeau, G.A.: Stdiff: Spatio-temporal diffusion for continuous stochastic video prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6666–6674 (2024)
39. Yuan, H., Zhang, S., Wang, X., Wei, Y., Feng, T., Pan, Y., Zhang, Y., Liu, Z., Albanie, S., Ni, D.: Instructvideo: Instructing video diffusion models with human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6463–6474 (2024)
40. Zha, K., Gao, Z., Shen, M., Hong, Z.W., Boning, D.S., Katabi, D.: RL tango: Reinforcing generator and verifier together for language reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
41. Zhao, M., Zhu, H., Xiang, C., Zheng, K., Li, C., Zhu, J.: Identifying and solving conditional image leakage in image-to-video diffusion model. In: Advances in Neural Information Processing Systems (2024)
42. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: DiffusionNFT: Online diffusion reinforcement with forward process. In: The Fourteenth International Conference on Learning Representations (2026)