Reward Lightning: Fast Video Generation via Homologous Preference Distillation

Jiaxiang Cheng*†, Bing Ma, Xuhua Ren, Kai Yu, Peng Zhang
Tianxiang Zheng, and Qinglin Lu

Tencent Hunyuan, China
jiaxiangcc@gmail.com

Abstract. Achieving simultaneous preference alignment and distillation acceleration in video diffusion models remains an open challenge. Existing methods optimize the two objectives over mismatched representation spaces, where improving one objective often compromises the other. To overcome this, we propose **Reward Lightning**, a unified framework that aligns and accelerates a video diffusion model within a single shared representation. Its central principle is *homology*: both objectives are evaluated on identical latent features, which mitigates the gradient conflicts that arise when they are optimized over disjoint representations. As a foundational component, we first introduce a latent reward model (LRM) that scores videos directly in the latent space, without decoding back to the pixel space. Building on the LRM, homologous preference distillation (HPD) reuses this shared backbone to perform adversarial distillation and preference alignment jointly, yielding few-step generators that remain faithful and well aligned. Extensive experiments demonstrate that the LRM surpasses pixel-level and latent-level reward baselines by 11.0% and 14.7% in preference accuracy, and that Reward Lightning generates high-fidelity videos in merely 1 to 4 steps, improving the average VBench score by 2.1% while leading in text alignment, motion quality, and visual quality. Project page: https://reward-lightning.github.io.

Keywords: Video Diffusion Models · Preference Alignment · Distillation Acceleration

1 Introduction

Video diffusion models [3, 12, 15, 24, 43, 47] have substantially advanced visual generation. Their large parameter counts, long temporal sequences, and iterative sampling, however, incur prohibitive inference costs. Distillation techniques [11, 27, 40, 44, 48, 57, 59, 62] effectively reduce the number of sampling steps, yet they inevitably distort the generative priors essential for reflecting human preferences [2, 25, 30, 46, 53, 54]. This raises a key question: how can we integrate preference alignment into the distillation trajectory without compromising fidelity?

*Corresponding Author. †Project Lead.

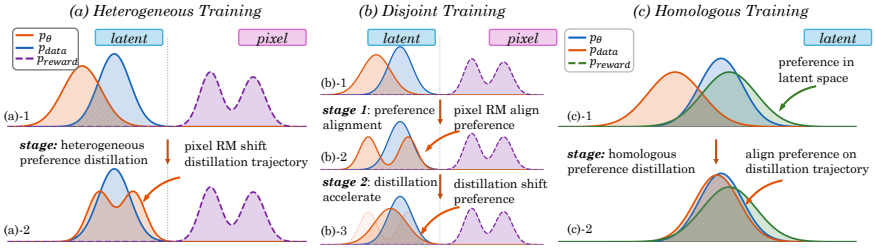

Fig. 1: Motivation. *(a) Heterogeneous Training*: Pixel rewards induce shifts in distillation trajectories through multi-objective gradient conflicts. *(b) Disjoint Training*: Sequential training discards original preference distributions through catastrophic forgetting. *(c) Homologous Training*: Preference distillation based on homologous structures and data guides the generator toward the jointly optimal distribution.

Recent methods such as DMDR [18], FlashDMD [5], and DOLLAR [8] integrate reinforcement learning with distillation to achieve aligned few-step generation. These paradigms, however, rely on heterogeneous optimization structures: they either linearly combine a pixel-space reward model with latent-space distillation, or update disjoint parameters for the two objectives (as shown in Fig. 1-(a,b)). In the context of multi-objective optimization, such structural and representational heterogeneity aggravates gradient conflicts, particularly in the high-noise regimes of few-step sampling, where reaching an optimal equilibrium between generation fidelity and human preference becomes difficult.

To overcome this bottleneck, we propose **Reward Lightning**, a unified framework built on the principle of *homology*, strict structural and representational consistency between preference alignment and distillation. Because both objectives are evaluated on identical latent features, homology removes the representational mismatch underlying the gradient conflicts above. Reward Lightning realizes this principle through two components: a Latent Reward Model (LRM) that supplies the shared backbone, and Homologous Preference Distillation (HPD) that learns over it. Recent work like PRFL [34] has elegantly shown that pre-trained video diffusion models are naturally suited to process-aware reward modeling in the latent space. We share this latent-evaluation philosophy, but design our LRM to play a dual role: not only as a standalone optimization target, but also as a structural prerequisite for preference distillation. To evaluate the highly blurred trajectory of few-step generation ($\leq$ 4 NFEs), we augment the preference dataset with few-step samples alongside their multi-step counterparts. Unlike existing methods [6, 27, 34] that apply time-step conditioning to the video representations alone while relying on a purely trainable query, we project the time-step embeddings directly as the dynamic query, allowing it to adaptively capture the varying representations along the flow trajectory.

The gradient conflict diagnosed above could in principle be resolved by explicit multi-objective optimization, but such treatment is computationally prohibitive for large-scale flow matching models. Building on the shared LRM back-

bone, HPD mitigates the conflict through a unified design rather than resolving it explicitly. By enforcing *structural homology* (a shared reward backbone) and *space homology* (the same latent representations), both the distillation and preference gradients are computed within an identical latent manifold. This homologous constraint acts as an implicit gradient regularizer: it promotes alignment between the gradient directions and stabilizes the joint optimization. In this way, the homology integrates preference alignment into the distillation trajectory, yielding few-step samples of both high fidelity and strong human alignment.

Extensive experiments validate the effectiveness of our unified framework. As a structural foundation, our LRM attains a preference accuracy on VideoGen-RewardBench [30] that surpasses existing pixel-level and latent-level baselines by 11.0% and 14.7%, respectively. Powered by this homology, our HPD reduces the sampling process to $1-4$ NFEs, as evaluated on VBench [16]. Under this few-step setting, our approach achieves a superior balance between inference efficiency and generation quality compared with both pure distillation and preference distillation methods. Specifically, it improves the average overall score by 2.1% and sets a new record in text alignment, motion quality, and visual quality.

2 Related Work

Post-Training. Post-training of diffusion models pursues two largely separate goals: preference alignment and distillation acceleration. To align outputs with human preferences, Reinforcement Learning from Human Feedback (RLHF) [1, 7, 36] has been adapted to diffusion models, either through learned reward models [23, 54] or Direct Preference Optimization (DPO) [37, 46, 55]. These alignment methods, however, typically operate on multi-step sampling [14, 45]. Distillation, by contrast, targets the high inference latency of iterative sampling. Progressive distillation [26, 33, 40], consistency models [20, 32, 44, 48], distribution matching [56, 57], and adversarial distillation [6, 27, 41, 42] all compress the generation trajectory into a few steps. Yet these methods prioritize fidelity preservation and distribution matching, and largely overlook human preference alignment. As a result, alignment and acceleration have advanced largely in isolation.

Joint Preference Optimization and Distillation. Recent methods [5, 8, 9, 18, 58] attempt to integrate preference optimization with distillation, so as to attain efficiency and alignment simultaneously [41, 46]. Sequential approaches address the two tasks in disjoint stages, but the second stage inevitably shifts the distribution learned in the first, a manifestation of catastrophic forgetting [10, 22]. Naive joint training, in turn, linearly combines a pixel-level preference reward with a latent-level distillation objective; this cross-domain mismatch induces severe gradient conflicts, as pixel-space optimization directly interferes with latent-space distillation. Neither paradigm maximizes both acceleration and alignment. The root cause is representational: rewards and distillation are optimized in mismatched spaces, which motivates aligning them within a single representation.

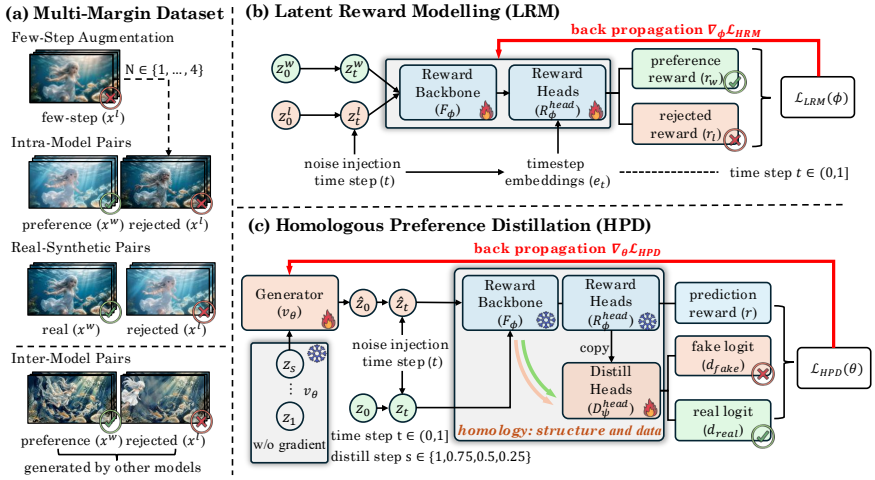

Fig. 2: Overall architecture of Reward Lightning. (a) Dataset: Fuzzy videos generated by $1-4$ NFEs as rejected videos to adapt to subsequent preference distillation. Other video pairs are used to improve the generalization of LRM. (b) LRM: Evaluates video pairs in latent space; the reward backbone inherits from a pre-trained model, while the reward head outputs video scores. We employ the Bradley-Terry with Ties (BTT) loss and introduce a regularization loss to prevent reward hacking. (c) HPD: The distilled head is extended from a reward head with identical architecture. Both heads interact with the same reward features in parallel, outputting rewards and logits. We provide dynamic preference weighting to ensure stability during initial training.

3 Methodology: Reward Lightning

In this section, we propose Reward Lightning, a unified framework for simultaneous preference alignment and distillation acceleration in flow matching models. Its central concept is homology: by enforcing structural and space homology, both the distillation and preference objectives are evaluated within a shared latent representation space. Below, we first review preference alignment and adversarial distillation (Sec. 3.1); then introduce our latent reward model, which serves as the structural foundation (Sec. 3.2); and finally detail the mechanism and optimization strategy of our framework (Sec. 3.3).

3.1 Preliminaries: Brief Review for Preference and Distillation

Reinforcement Learning from Human Feedback. Reinforcement Learning from Human Feedback (RLHF) [2, 23, 25, 52, 54] aligns a pre-trained generator $v_\theta(\cdot, t)$ with human preferences, typically through a pixel-space reward model $R_\phi(\cdot)$. Given a preference dataset $\mathcal{D}_{\text{pair}} = \{(\mathbf{x}^{\text{w}}, \mathbf{x}^{\text{l}})\}$ comprising preferred videos $\mathbf{x}^{\text{w}}$ and rejected videos $\mathbf{x}^{\text{l}}$, $R_\phi(\cdot)$ is optimized via the Bradley-Terry (BT) objective [4]:

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(\mathbf{x}^{\text{w}},\mathbf{x}^{\text{l}})\sim\mathcal{D}_{\text{pair}}}\left[\log\sigma\left(R_\phi(\mathbf{x}^{\text{w}}) - R_\phi(\mathbf{x}^{\text{l}})\right)\right], \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. In standard latent-based generative frameworks [39], the optimization occurs in the latent space $\mathbf{z} \in \mathcal{Z}$, where $\mathbf{z} = g^E_{\text{VAE}}(\mathbf{x})$ is extracted via a pre-trained Variational Auto-Encoder (VAE) encoder [21]. To transfer human preferences to the generator $v_\theta(\cdot, t)$, Reward Feedback Learning (ReFL) [54] directly maximizes the reward of the predicted clean flow state $\hat{\mathbf{z}}_0^{\text{w,l}} = f_\theta(\mathbf{z}_t^{\text{w,l}}, t)$, where $f_\theta(\mathbf{z}_t, t) := \mathbf{z}_t - t \cdot v_\theta(\mathbf{z}_t, t)$ represents the flow trajectory mapping toward $t = 0$. Because the reward model evaluates pixel-space representations, the ReFL objective is formulated with an explicit decoding step:

$$\mathcal{L}_{\text{ReFL}}(\theta) = -\mathbb{E}_{t\in(0,1]}\left[R_\phi(g^D_{\text{VAE}}(f_\theta(\mathbf{z}_t, t)))\right]. \tag{2}$$

Adversarial Distillation. Adversarial distillation accelerates iterative sampling by compelling the generator $v_\theta(\cdot, t)$ to map the noise prior $\mathbf{z}_1 \sim \mathcal{N}(0; \mathbf{I})$ directly to the real latent data distribution $p_{\text{data}}(\mathbf{z})$ in a few steps. It introduces a trainable discriminator $D_\psi(\cdot, \cdot)$ to distinguish between real latent samples $\mathbf{z}_0$ and generated predictions $\hat{\mathbf{z}}_0$. To evaluate representations across varying noise levels, intermediate states $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$ are constructed by perturbing $\mathbf{z}_0$ and $\hat{\mathbf{z}}_0$ with noise at timestep t. Typically, the discriminator is trained via the standard Hinge loss [35]:

$$\mathcal{L}^D_{\text{ADV}}(\psi) = \frac{1}{2}\mathbb{E}_{\mathbf{z}_0\sim p_{\text{data}}}\left[\max(0, 1 - D_\psi(\mathbf{z}_t, t))\right] + \frac{1}{2}\mathbb{E}_{\hat{\mathbf{z}}_0\sim p_\theta}\left[\max(0, 1 + D_\psi(\hat{\mathbf{z}}_t, t))\right]. \tag{3}$$

The generator is optimized to deceive the discriminator via the adversarial loss:

$$\mathcal{L}^G_{\text{ADV}}(\theta) = -\mathbb{E}_{\hat{\mathbf{z}}_0\sim p_\theta}\left[D_\psi(\hat{\mathbf{z}}_t, t)\right]. \tag{4}$$

This minimax formulation ensures that the generator produces high-fidelity structural distributions without relying on multi-step solver evaluations.

3.2 Latent Reward Model: As a Structural Prerequisite

Multi-Margin Dataset Construction. Constructing a generalizable reward model requires a dataset that spans diverse distributions. As shown in Fig. 2-a, we curate a multi-margin dataset comprising preferred and rejected videos across distinct comparative scenarios (more annotation details in the appendix):

- *Intra-Model Pairs.* We collect 30,000 pairs generated by the target model across random seeds. Five professional annotators established preferences based on visual quality, motion dynamics, and semantic alignment. This subset drives self-refinement, steering the generator toward its optimal modes.
- *Inter-Model Pairs.* Sourced from VisionRewardDB [53], this subset contains 17,390 expertly annotated pairs generated by five state-of-the-art models. This subset prevents the reward model from overfitting to internal distributions, and enforces robust boundary generalization.
- *Real-Synthetic Pairs.* We construct 12,000 pairs comparing real-world videos with generated videos. We designate the real videos as preferred and the generated videos as rejected. These pairs raise the generation quality ceiling.

Few-step Augmentation for Distillation. Standard preference optimization evaluates high-fidelity multi-step samples. However, preference distillation entails evaluating accelerated few-step predictions spanning $N \in \{1, 2, 3, 4\}$ steps that often exhibit severe structural blurring and thus act as out-of-distribution data. To mitigate this domain gap, we augment the dataset with $5,000$ specialized pairs. We pair high-quality multi-step outputs as preferred targets with their structurally degraded few-step counterparts as rejected samples. This targeted augmentation equips the reward model to provide reliable gradient guidance across highly truncated distillation trajectories.

Reward Model Architecture. As shown in Fig. 2-b, we initialize our LRM backbone $F_\phi(\cdot, t)$ with a pre-trained diffusion model [47]. Given a preference latent pair $(\mathbf{z}_0^{\mathrm{w}}, \mathbf{z}_0^{\mathrm{l}})$, we sample intermediate flow states $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$, where $t \sim \mathcal{U}(0,1)$ and $\mathbf{z}_1 \sim \mathcal{N}(0; \mathbf{I})$. The backbone extracts spatial features $F_t = F_\phi(\mathbf{z}_t, t)$. This latent-space evaluation bypasses VAE decoding, ensuring preference and distillation gradients share an identical manifold.

Time-Modulated Attention Head. To evaluate representations across the severe distributional shifts of the flow trajectory, the reward head $R_\phi^{\mathrm{head}}(\cdot, \cdot)$ must be noise-aware. Unlike traditional architectures that rely solely on learnable queries and struggle to generalize across varying timesteps [6, 27, 34], we dynamically modulate the continuous flow state. Specifically, for each of the H=3 attention heads h, we project the timestep embedding $\mathbf{e}_t$ via a weight matrix W_h into a dynamic state query $\mathbf{s}_{t,h}$. This time-dependent query adaptively aggregates the spatial features F_t via cross-attention. A linear projection W_o then maps the concatenated outputs to the final scalar reward:

$$R_\phi^{\mathrm{head}}(F_t, t) = W_o \, \mathrm{Concat}_{h=1}^{H} \big(\mathrm{CrossAttn}(\mathbf{s}_{t,h}, F_t) \big). \tag{5}$$

This temporal modulation allows the head to shift its receptive focus—capturing macroscopic structural layouts in high-noise regimes and fine-grained textures in low-noise regimes—providing precise, state-aware gradient guidance.

Reward Model Training We align our LRM with human preferences using the Bradley-Terry-with-Ties (BTT) objective [30], an extension of BT that additionally models tie outcomes (full formulation in the appendix). While generally effective, applying the BTT loss to our multi-margin dataset exposes a critical vulnerability. The BTT loss continuously maximizes reward differences; consequently, for large-margin pairs (e.g., real vs. synthetic videos), the model tends to exploit superficial artifacts for trivial domain separation. This shortcut learning causes feature collapse, degrading the sensitivity of the model to the fine-grained visual differences critical for narrow-margin pairs (e.g., intra-model pairs). To mitigate this issue, we propose a dynamic margin-clipping mechanism guided by an Exponential Moving Average (EMA). During training, each batch is partitioned into a narrow-margin subset $\mathcal{B}_{\mathrm{narrow}}$ and a large-margin subset $\mathcal{B}_{\mathrm{large}}$. Let the predicted reward difference for a given pair be

$\Delta r = R_\phi^{\text{head}}(F_t^{\text{w}}, t) - R_\phi^{\text{head}}(F_t^{\text{l}}, t)$. We track the EMA of the reward differences exclusively on the narrow-margin subset:

$$\mu \leftarrow \rho\mu + (1-\rho)\frac{1}{|\mathcal{B}_{\text{narrow}}|}\sum_{i\in\mathcal{B}_{\text{narrow}}} \Delta r_i, \tag{6}$$

where $\rho = 0.95$ is the decay factor. This dynamic scalar μ serves as an adaptive threshold to bound the loss. The reward difference for any large-margin pair $j \in \mathcal{B}_{\text{large}}$ is clipped by μ. The overall training objective is formulated as:

$$\mathcal{L}_{\text{LRM}}(\phi) = -\mathbb{E}_{i\in\mathcal{B}_{\text{narrow}}}[\log\sigma(\Delta r_i)] - \mathbb{E}_{j\in\mathcal{B}_{\text{large}}}[\log\sigma(\min(\Delta r_j, \mu))]. \tag{7}$$

This adaptive clipping acts as an implicit regularizer against reward hacking. When the reward margin of easy samples exceeds the dynamic threshold established by hard samples, their gradient contribution is zeroed out. Consequently, the optimization is compelled to focus on fine-grained human alignment rather than exploiting superficial domain discrepancies.

3.3 Homologous Preference Distillation: Faster and Better

Key Motivation Analysis: Mitigating Multi-Objective Conflicts. We formulate the joint training of adversarial distillation and preference alignment on flow states $\mathbf{z}_t \in \mathcal{Z}$ as a multi-objective optimization problem. The distillation objective $\mathcal{L}_{\text{ADV}}^G$ focuses on data distribution matching, whereas the preference objective $\mathcal{L}_{\text{ReFL}}$ targets specific human preference. This task divergence naturally induces gradient conflicts. In traditional heterogeneous paradigms, such as cross-space pixel rewards or disjoint architectures, the gradients for these two objectives are computed over unaligned representation manifolds. Lacking a shared feature basis to constrain the optimization, the multi-objective conflict is exacerbated, often leading to out-of-distribution drift during generation.

Although our approach remains a multi-objective framework, it effectively mitigates these inherent conflicts via strict structural and data alignment. By evaluating the identical intermediate flow state $\mathbf{z}_t$ through a shared feature backbone F_ϕ, both tasks are anchored to a unified representation space $F_t = F_\phi(\mathbf{z}_t, t)$. Consequently, the optimization gradients, $\nabla_{\mathbf{z}_t}\mathcal{L}_{\text{ADV}}^G$ and $\nabla_{\mathbf{z}_t}\mathcal{L}_{\text{ReFL}}$, are constrained by the exact same semantic basis. This homologous formulation acts as an implicit gradient regularizer. Instead of explicitly eliminating the objective divergence, it restricts the optimization directions within a shared manifold, promoting gradient alignment. This mechanism ensures that the generator stably converges to the optimal joint support of visual fidelity and human preference, bypassing the need for computationally expensive explicit gradient matching.

Homologous Model Design To actualize the homologous mechanism, we design a modular dual-head architecture. Our LRM serves as a shared backbone for latent feature extraction, branching into two parallel heads to support simultaneous preference alignment and adversarial distillation: a frozen reward head

Algorithm 1 Homologous Preference Distillation

Require: Generator v_θ, Reward backbone F_ϕ, Reward head R_ϕ^{head}, Dataset $\mathcal{D}_{\text{train}}$
Initialize: Discriminator D_ψ^{head} weights $\leftarrow R_\phi^{\text{head}}$
while not converged **do**
  // **Distillation Sampling**
  Sample $\mathbf{z}_1 \sim \mathcal{N}(0;\mathbf{I})$, $\{\mathbf{z}_0, c\} \sim \mathcal{D}_{\text{train}}$ and $s \in \{1, 0.75, 0.5, 0.25\}$
  Compute: $\mathbf{z}_s = \text{sg}\left(\text{ODESolver}(v_\theta, \mathbf{z}_1, c, 1 \to s)\right)$
  Compute: $\hat{\mathbf{z}}_0 = \mathbf{z}_s - s \cdot v_\theta(\mathbf{z}_s, c)$
  // **Construct Homologous States**
  Sample $\mathbf{z}_1 \sim \mathcal{N}(0;\mathbf{I})$ and $t \sim \mathcal{U}[0,1]$
  Compute: $\hat{\mathbf{z}}_t = (1-t)\hat{\mathbf{z}}_0 + t\mathbf{z}_1; \quad \mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$
  // **Discriminator Optimization**
  Compute: $F_t^{\text{fake}} = F_\phi(\hat{\mathbf{z}}_t, t); \quad F_t^{\text{real}} = F_\phi(\mathbf{z}_t, t)$
  Compute: $\mathcal{L}_{\text{ADV}}^D = \frac{1}{2}(\max(0, 1 - D_\psi^{\text{head}}(F_{\text{real}}, t)) + \max(0, 1 + D_\psi^{\text{head}}(F_{\text{fake}}, t)))$
  Update: $\psi \leftarrow \psi - \eta_D \nabla_\psi \mathcal{L}_{\text{ADV}}^D$
  // **Generator Optimization: Adversarial and ReFL loss**
  Compute: $\mathcal{L}_{\text{ReFL}} = -R_\phi^{\text{head}}(F_{\text{fake}}, t)$
  Compute: $\mathcal{L}_{\text{ADV}}^G = -D_\psi^{\text{head}}(F_{\text{fake}}, t)$
  Compute: $\lambda = \exp\left(-\text{ReLU}\left(\text{sg}\left(\mathcal{L}_{\text{ADV}}^D\right) - 0.5\right)^2\right)$
  Compute: $\mathcal{L}_{\text{HPD}} = \mathcal{L}_{\text{ADV}}^G - \lambda \cdot \mathcal{L}_{\text{ReFL}}$
  Update: $\theta \leftarrow \theta - \eta_G \nabla_\theta \mathcal{L}_{\text{HPD}}$
end while

$R_\phi^{\text{head}}(\cdot, t)$ and a trainable discriminator head $D_\psi^{\text{head}}(\cdot, t)$. To ensure structural simplicity, we initialize the discriminator D_ψ^{head} directly from the weights of the reward head R_ϕ^{head}. This direct weight inheritance provides a robust semantic inductive bias. By transferring the comprehensive visual understanding of the reward model to the discriminator, this unified design stabilizes the early phase of adversarial training and accelerates convergence.

Unified Feature Routing. As shown in Fig. 2-c, the forward pass operates as a synchronized evaluation pipeline. The generator v_θ first predicts a velocity vector to reconstruct the clean latent representation $\hat{\mathbf{z}}_0$. Because the shared backbone is pre-trained on intermediate flow states, we construct the corresponding flow state via linear interpolation: $\hat{\mathbf{z}}_t = (1-t)\hat{\mathbf{z}}_0 + t\mathbf{z}_1$, where $\mathbf{z}_1 \sim \mathcal{N}(0;\mathbf{I})$. The shared backbone then extracts the latent features $F_t = F_\phi(\hat{\mathbf{z}}_t, t)$. We route these features in parallel to compute the preference score $r = R_\phi^{\text{head}}(F_t, t)$ and the adversarial logit $d = D_\psi^{\text{head}}(F_t, t)$. Evaluating both tasks on identical features strictly enforces representational alignment, ensuring that the gradient updates for distillation and preference are anchored to the exact same semantic domain.

Policy Distillation Objective. We optimize the generator v_θ and a trainable discriminator head D_ψ^{head}, while keeping the shared backbone F_ϕ and the reward head R_ϕ^{head} frozen. The discriminator head is updated via the Hinge adversarial loss, denoted as $\mathcal{L}_{\text{ADV}}^D$. Concurrently, the generator minimizes the adversarial

loss $\mathcal{L}_{\text{ADV}}^{G}$ to align with the real data distribution and maximizes the preference score $\mathcal{L}_{\text{ReFL}}(\theta) = -\mathbb{E}[R_{\phi}^{\text{head}}(F_t, t)]$ provided by the frozen reward head. However, a direct linear combination of these objectives causes severe optimization instability. During the early training phase, the generated states deviate from the target data distribution. Enforcing alignment thus produces unreliable gradients.

Adaptive Preference Weight. To ensure stable convergence, we introduce an adaptive preference weighting mechanism. The discriminator loss is high for easily distinguishable artifacts, and approaches a Nash equilibrium value of 0.5 for realistic samples. We project the deviation from this equilibrium into a target preference weight λ using a half-Gaussian squash function:

$$\hat{\lambda} = \exp\left(-\text{ReLU}\left(\text{sg}\left(\mathcal{L}_{\text{ADV}}^{\text{D}}\right) - 0.5\right)^2\right). \tag{8}$$

This continuous mapping operates as a soft bounding constraint. When the generated state diverges from the target data distribution ($\mathcal{L}_{\text{ADV}}^{\text{D}} \to \infty$), the exponential function heavily penalizes the target weight ($\hat{\lambda} \to 0$). This step mathematically suppresses the preference gradient. As structural realism improves, the loss approaches the equilibrium ($\mathcal{L}_{\text{ADV}}^{\text{D}} \to 0.5$). The penalty vanishes, smoothly activating the preference objective ($\hat{\lambda} \to 1$). To defend against erratic spikes in the discriminator loss, we apply an exponential moving average. This operation decouples the dynamic weight from single-step fluctuations:

$$\lambda = \beta \cdot \lambda^{-} + (1 - \beta) \cdot \hat{\lambda}, \tag{9}$$

where $\beta = 0.99$ acts as the momentum factor. This dual constraint mathematically guarantees the safe and gradual introduction of the preference gradient. The preference gradient activates only when the generator exhibits sustained realism. The total training objective for the generator is formulated as:

$$\mathcal{L}_{\text{HPD}}(\theta) = \mathcal{L}_{\text{ADV}}^{G} + \lambda \cdot \mathcal{L}_{\text{ReFL}}. \tag{10}$$

By dynamically coupling these objectives, this formulation acts as a self-paced curriculum. It safely anchors the optimization trajectory and ensures the stable convergence of the generator into the optimal joint distributional support.

4 Experiments

4.1 LRM: Reward Learning

Datasets & Training Settings. *Datasets.* We train the LRM using the multi-margin dataset described in Sec. 3.2. All captions and prompts are from Koala-36M [50]. We provide comprehensive statistics and human annotation details in the appendix. *Training Settings.* We employ Wan2.2-14B [47] as the base architecture for all reward modeling experiments. The reward model is optimized using AdamW [31] with betas of 0.9 and 0.999, a weight decay of 1×10^{-3}, and

Table 1: Quantitative comparison of preference accuracy and online reward feedback efficiency. We report preference accuracy on out-of-distribution benchmarks, together with the training latency, memory consumption, and throughput of online reward feedback. All computations run on 8× H20 GPUs. Underline: The second-best performance. **Bold**: Best performance.

Method	VideoGen-RewardBench		GenAI-Bench		Reward Feedback Efficiency			
	Overall Accuracy (↑)		Overall Accuracy (↑)		Resolution	Frames	Train Memory	Train Latency
	w/ Ties	w/o Ties	w/ Ties	w/o Ties				
Pixel Reward Models								
Random (QwenVL-2 [49])	41.86	50.30	33.67	49.84	480 × 480	≤ 1	-	-
LiFT [51]	39.08	57.26	37.06	58.39	384 × 384	≤ 1	87.81	213.2
VideoScore [13]	41.80	50.22	49.03	71.69	384 × 384	≤ 1	85.60	207.4
VisionRewrd [53]	56.77	67.59	51.56	72.41	224 × 224	≤ 1	82.15	185.3
VideoAlign [30]	<u>61.26</u>	<u>73.59</u>	49.41	<u>72.89</u>	448 × 448	≤ 1	87.46	217.9
Latent Reward Models								
Random (Wan2.2 [47])	48.41	58.03	40.12	53.88	1280 × 720	≥ 81	-	-
PRFL [34]	57.59	69.20	<u>53.40</u>	72.10	1280 × 720	≥ 81	77.32	156.8
Ours	**72.24**	**86.63**	**59.85**	**78.44**	1280 × 720	≥ 81	76.24	154.2

a constant learning rate of 1×10^{-5}. We process the 5-second paired videos at a resolution of 720P. For memory management, we apply three memory optimization strategies. (1) Offload: we transfer frozen models (e.g., VAE [21] and text encoder [38]) between CPUs and GPUs, which reduces the peak memory before backpropagation. (2) Fully Sharded Data Parallel (FSDP) [60]: sharding model parameters, gradients, and optimizer states across GPUs; (3) Ulysses Sequence Parallelism (SP) [17]: partitioning latent video tokens within self-attention blocks to enable efficient processing of long temporal sequences.

Benchmarks & Evaluation Metrics. To evaluate the human alignment and generalization of the reward model, we employ two popular out-of-distribution (OOD) video reward benchmarks. (1) VideoGen-RewardBench [30]: This benchmark generates video pairs based on prompts from VideoGen-Eval, which includes different modern T2V models. (2) GenAI-Bench [19]: It generates short video pairs using early T2V models, providing a low-quality distribution for evaluating generalization. Given the pairwise preference labels, we adopt preference accuracy as our evaluation metric, which directly reflects the agreement between model predictions and human annotations. We compare LRM against baselines including pixel reward models [13, 30, 51, 53] and latent ones [34].

Main Results. Tab. 1 demonstrates that our LRM achieves state-of-the-art preference accuracy on VideoGen-RewardBench [30] and GenAI-Bench [19], reaching 72.24% and 59.85%, respectively, surpassing VideoAlign [30], the strongest pixel-level baseline, by over 10%. These results reflect precise alignment with human judgment across three core dimensions: Text Alignment, comprising subject and background consistency; Motion Quality, involving smoothness and dynamic degree; and Visual Quality, which includes aesthetic and image quality. LRM also

Table 2: Quantitative comparison of Text-to-Video (T2V) and Image-to-Video (I2V) on VBench. For a fair comparison, we build all models on the Wan2.2 architecture. With the exception of TurboDiffusion, we reproduce all results ourselves. VQ, MQ, and TA denote Visual Quality, Motion Quality, and Text Alignment.

Methods	Training Type	VBench				VBench-I2V			
		TA (↑)	MQ (↑)	VQ (↑)	Quality Score (↑)	TA (↑)	MQ (↑)	VQ (↑)	I2V Score (↑)
80-NFEs									
Wan2.2 [47]	Pretrain	97.34	79.59	69.48	84.23	95.87	74.64	67.63	92.91
RGB ReFL [28]	PixelRL	97.45	80.12	69.55	84.41	96.12	75.24	67.85	93.52
PRFL [34]	LatentRL	97.83	81.56	70.45	85.72	96.65	76.58	68.61	94.87
4-NFEs									
Wan2.2 [47]	Pretrain	80.15	61.24	56.43	68.52	80.41	61.81	55.94	75.97
TurboDiffusion [59]	Distill	94.53	75.62	63.28	83.51	95.84	74.25	64.87	92.45
DMD2 [56]	Distill	91.46	72.53	64.38	81.27	94.28	73.33	67.78	91.92
DMDR [18]	PixelRL & Distill	96.14	71.85	68.57	84.63	96.17	72.14	69.56	93.28
FlashDMD [5]	PixelRL & Distill	88.19	68.23	67.84	82.36	92.15	69.45	69.12	92.59
HPD	LatentRL & Distill	**97.26**	**77.48**	**68.71**	**85.83**	**96.32**	**75.39**	**69.78**	**95.34**
1-NFEs									
Wan2.2 [47]	Pretrain	70.45	53.27	51.18	61.26	72.84	55.73	51.89	69.59
APT [27]	Distill	85.37	65.42	**64.19**	78.53	87.89	68.18	61.81	84.87
HPD	LatentRL & Distill	**92.54**	**70.16**	62.83	**82.42**	**95.14**	**73.39**	**65.31**	**91.92**

improves training efficiency by operating directly in the latent space rather than the pixel space. By avoiding expensive VAE decoding and high-dimensional pixel convolutions, LRM lowers both training memory and latency at 720P.

Ablation Study. Tab. 3-(a-c) highlights the contribution of each component to LRM performance. The dataset composition study in part (a) establishes that the full multi-margin collection is indispensable for preference alignment, as excluding few-step augmentation or real-synthetic pairs leads to a significant decline in accuracy and reward scores. In terms of head architecture, time-step projection in part (b) proves superior to self-attention and trainable query baselines. This design allows the model to adaptively extract robust features from the latent space z by directly conditioning on the temporal embedding across the flow trajectory. Reward regularization in part (c) is also critical for preventing reward hacking, as it mathematically bounds reward gaps for stable feedback.

4.2 HPD: Improving and Accelerating Diffusion Models

Datasets & Training Settings. Unlike the paired video datasets, we train the generator on single captioned videos. We collect a high-quality real-video dataset comprising Koala-36M [50] and Intern4K [29]. We build our generator on Wan2.2-14B [47], and initialize the discriminator from LRM. In the training process, we employ AdamW [31] with betas of 0.5 and 0.999 for lower memory consumption. This optimizer adopts a weight decay of 1×10^{-3} and a constant learning rate of 5×10^{-7}. The generator is trained with the batch size of 64 and 2,000 steps. An EMA of 0.995 is applied to the generator. For memory efficiency,

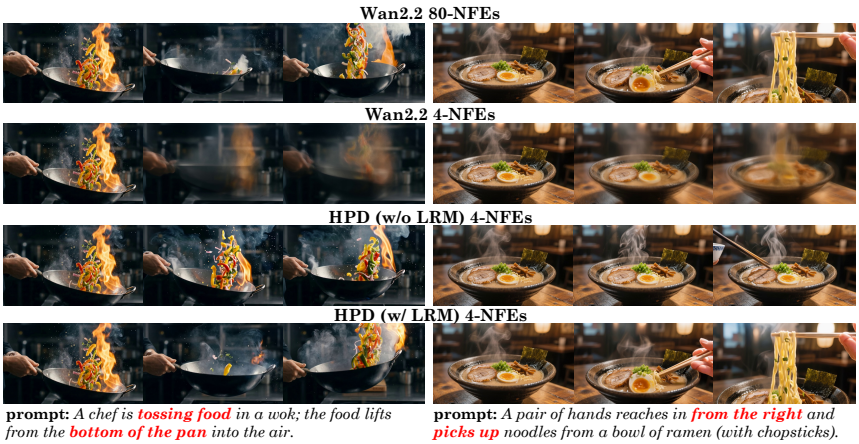

Fig. 3: Qualitative Results. Upper: Visualization of 80-NFEs and 4-NFEs from the Wan2.2-I2V-A14B. Bottom: Built on the baseline model, compared against a distillation without homologous preference distillation.

we extend the offloading strategy to the EMA generator in addition to the frozen VAE and text encoder. Furthermore, we employ a full-sharding strategy via FSDP and SP across the generator, reward model, and discriminator to enable stable training within the limited hardware constraints.

Benchmarks & Evaluation Metrics. We evaluate the generator on VBench [16] for comprehensive video assessment. The assessment involves multi-dimensional metrics, specifically Visual Quality, Motion Quality, and Text Alignment. To quantify the distribution gap between generated and real-world videos, we calculate the Fréchet Video Distance (FVD). Additionally, we conduct a human evaluation based on voting results to capture subjective visual preferences.

Baseline Models. We evaluate HPD against state-of-the-art baselines in three categories: heterogeneous training, standalone distillation, and preference alignment. For heterogeneous paradigms, we include DMDR [18] and FlashDMD [5], which represent joint optimization using disjoint representation spaces. The distillation-only baselines comprise TurboDiffusion [59] and DMD2 [56] to represent current sampling acceleration techniques. For preference alignment, we compare against PRFL [34], a multi-step reward feedback method.

Main Results. Tab. 2 summarizes the quantitative performance of HPD under the 4-NFE and 1-NFE few-step settings on both VBench (T2V) and VBench-I2V [16]. Compared to both pure distillation and heterogeneous training paradigms, our approach achieves a superior balance between inference efficiency and generation quality. In the 4-NFE setting, our framework attains the highest overall qual-

Fig. 4: Qualitative Results For fair comparison, all models are built on Wan2.2-I2V-A14B. Except for TurboDiffusion, we reproduce all results ourselves. *Row* 1: ReFL model based on latent reward model. *Row* $2-3$: Heterogeneous model, composed of pixel ReFL and DMD distillation. *Row* 4: Distillation model, primarily based on rCM method. *Row* 5: Homologous model, with structural and space homology.

ity score among heterogeneous methods such as DMDR [18] and FlashDMD [5], leading across text alignment, motion quality, and visual quality. This indicates that homologous alignment effectively mitigates the multi-objective conflicts that often degrade performance in disjoint architectures. While pure distillation baselines like TurboDiffusion [59] and DMD2 [56] focus primarily on acceleration, they lack the semantic guidance provided by human preference models. Our approach addresses this by integrating preference alignment directly into the distillation trajectory, yielding higher scores across three core dimensions: Text Alignment, Motion Quality, and Visual Quality. Even at the extreme limit of 1-NFE, our approach consistently surpasses the APT baseline in text alignment and motion quality, and attains the best overall quality score. As shown in Figs. 3 to 4, HPD eliminates the motion blur and temporal flickering prevalent in few-step sampling, producing stable motion and crisp visual details. The human evaluation in Fig. 5 corroborates these gains, with annotators favoring our results over competing baselines.

Ablation Study. Tab. 3-(d-e) underscores the necessity of structural homology and adaptive weighting within HPD. Initializing the discriminator with LRM parameters in part (d) provides a consistent semantic prior that raises the quality

Table 3: Ablation study on key components. (a)-(c) LRM: We provide ablations on the dataset composition, head architecture, and reward loss functions. (d)-(e) HPD: We ablate the distilled parameters and preference-loss weighting.

(a) Dataset Composition

Setting	Acc. (↑)	Score (↑)
Full	72.24	95.34
w/o Few-Step	68.47	93.81
w/o Real-Syn.	65.12	92.45

(b) Head Architecture

Setting	Acc. (↑)	Score (↑)
Self-Attention	64.38	91.87
w/ Query	69.85	94.12
w/ Time Proj.	72.24	95.34

(c) Reward Model Loss

Setting	Acc. (↑)	Score (↑)
w/o reg.	62.15	90.76
$\rho = 0.5$	68.91	93.58
$\rho = 0.95$	72.24	95.34

(d) Head Parameters

Setting	TA (↑)	MQ (↑)	VQ (↑)	Score (↑)
Random	91.14	68.58	62.87	88.26
Pre-trained	94.57	73.12	66.74	93.48
LRM	96.32	75.39	69.78	95.34

(e) Adaptive Preference Weight

Setting	TA (↑)	MQ (↑)	VQ (↑)	Score (↑)
$\lambda = 0$	91.45	71.34	67.12	90.87
$\lambda = 1$	93.12	72.18	67.58	92.73
$\lambda \propto -\mathcal{L}_{\text{ADV}}^{D}$	96.32	75.39	69.78	95.34

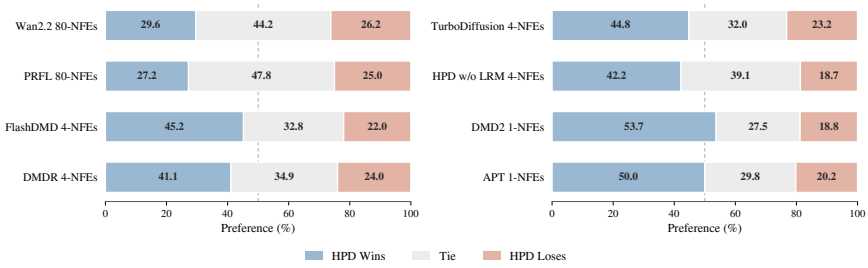

Fig. 5: Human Evaluation. We report the Win/Tie/Lose rate for HPD $1, 4-$NFEs, compared with the Euler baseline, distillation, and heterogeneous models.

ceiling and accelerates convergence compared to random baselines. In part (e), the adaptive weight λ ensures training stability by delaying preference alignment until the generator achieves structural realism. This self-paced mechanism prevents early gradient interference and out-of-distribution drift, guiding the model toward the optimal joint support of visual fidelity and human preference.

4.3 Further Analysis

Gradient Conflict Analysis. To validate that homology mitigates these conflicts, we measure the cosine similarity between the adversarial gradient $\nabla_\theta \mathcal{L}_{\text{ADV}}^{G}$ and the preference gradient $\nabla_\theta \mathcal{L}_{\text{ReFL}}$, averaged over 1,000 training iterations (Tab. 4-a). In heterogeneous paradigms with a pixel-space or separate-backbone reward model, the two gradients exhibit negative cosine similarity (-0.14 and -0.06), confirming that the optimization directions actively conflict. By contrast, HPD evaluates both objectives through a shared backbone on identical latent representations, raising the cosine similarity to 0.41. This confirms that structural and space homology acts as an implicit gradient regularizer.

Table 4: Further analysis on the homologous design. (a) Gradient conflict analysis: cosine similarity between $\nabla_\theta \mathcal{L}^G_{\text{ADV}}$ and $\nabla_\theta \mathcal{L}_{\text{ReFL}}$, averaged over 1,000 iterations (4-NFE, I2V); positive values indicate aligned gradients. **(b)** LRM robustness and the impact of HPD homology across distillation paradigms.

(a) Gradient Conflict

Setting	Backbone	Cosine Similarity (↑)
Pixel Reward + Distillation	Separate	−0.14
Latent Reward + Distillation	Separate	−0.06
HPD (Ours)	Shared	**0.41**

(b) LRM Robustness

Methods	TA (↑)	MQ (↑)	VQ (↑)	I2V Score (↑)	FVD (↓)
rCM [61]	95.84	74.25	64.87	93.45	207.32
+ LRM	96.65 +0.81	74.98 +0.73	65.82 +0.95	94.32 +0.87	205.45 -1.87
DMD2 [56]	94.28	73.33	67.78	91.92	218.78
+ LRM	95.07 +0.79	74.31 +0.98	68.63 +0.85	92.71 +0.79	217.64 -1.14
HPD	94.14	72.92	67.44	93.09	167.14
+ LRM	96.32 +2.18	75.39 +2.47	69.78 +2.34	95.34 +2.25	162.87 -4.27

Efficiency & Quality Analysis. As shown in Fig. 6, across 1 to 8 sampling steps, HPD maintains a superior efficiency-quality profile over DMD2 [56], TurboDiffusion [59], DMDR [18], and FlashDMD [5]. Unlike baselines that suffer structural drift, our homologous design anchors the trajectory within a unified manifold, delivering higher fidelity at lower cost. Representational homology thus enables near-teacher performance at 1-NFE, closing the speed–alignment gap. Notably, this advantage widens as the step count decreases, indicating that representational consistency matters most in the aggressive few-step regime.

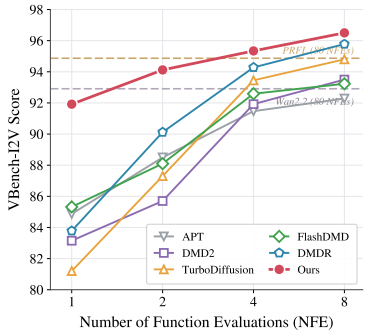

Fig. 6: Efficiency-quality Pareto front on Wan2.2-I2V-14B. Dotted lines denote 80-NFE baselines. HPD dominates the top-left region across 1 to 8 NFEs.

Distillation Generalizability Analysis. As shown in Tab. 4-b, LRM yields consistent gains across paradigms like rCM [61] and DMD2 [56]. However, the largest gains arise within HPD, where homology minimizes gradient conflicts and remains vital for the optimal ceiling in acceleration and human alignment.

5 Conclusion

We have introduced Reward Lightning, a unified framework that aligns representations through structural and space homology for joint preference alignment and distillation acceleration. Evaluating both objectives within an identical manifold minimizes the angle between conflicting gradients, acting as an implicit regularizer that drives the generator toward the joint support of high fidelity and human alignment. It accelerates inference well beyond heterogeneous methods.

References

1. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)
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), https://openreview.net/forum?id=YCWjhGrJFD
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
4. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952), http://www.jstor.org/stable/2334029
5. Chen, G., Huang, S., Liu, K., Zhu, J., Qu, X., Chen, P., Cheng, Y., Sun, Y.: Flash-dmd: Towards high-fidelity few-step image generation with efficient distillation and joint reinforcement learning. arXiv preprint arXiv:2511.20549 (2025)
6. Cheng, J., Ma, B., Ren, X., Jin, H.H., Yu, K., Zhang, P., Li, W., Zhou, Y., Zheng, T., Lu, Q.: Phased one-step adversarial equilibrium for video diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 3237–3245 (2026)
7. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
8. Ding, Z., Jin, C., Liu, D., Zheng, H., Singh, K.K., Zhang, Q., Kang, Y., Lin, Z., Liu, Y.: Dollar: Few-step video generation via distillation and latent reward optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17961–17971 (2025)
9. Eyring, L., Karthik, S., Dosovitskiy, A., Ruiz, N., Akata, Z.: Noise hypernetworks: Amortizing test-time compute in diffusion models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=DbzREoPwmM
10. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999)
11. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=uWj4s7rMnR
12. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=Fx2SbBgcte
13. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 2105–2123 (2024)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
16. 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)
17. Jacobs, S.A., Tanaka, M., Zhang, C., Zhang, M., Song, S.L., Rajbhandari, S., He, Y.: Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. arXiv preprint arXiv:2309.14509 (2023)
18. Jiang, D., Liu, D., Wang, Z., Wu, Q., Li, L., Li, H., Jin, X., Liu, D., Li, Z., Zhang, B., et al.: Distribution matching distillation meets reinforcement learning. arXiv preprint arXiv:2511.13649 (2025)
19. Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., Chen, W.: Genai arena: An open evaluation platform for generative models. Advances in Neural Information Processing Systems **37**, 79889–79908 (2024)
20. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=ymjI8feDTD
21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
22. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
23. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems **36**, 36652–36663 (2023)
24. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
25. 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)
26. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
27. Lin, S., Xia, X., Ren, Y., Yang, C., Xiao, X., Jiang, L.: Diffusion adversarial post-training for one-step video generation. In: International Conference on Machine Learning. pp. 37959–37974. PMLR (2025)
28. Lin, W., Chen, R., Liu, B., Yan, S., Feng, R., Wei, J., Zhang, Y., Zhou, Y., Feng, C., Ran, J., et al.: Contentv: Efficient training of video generation models with limited compute. arXiv preprint arXiv:2506.05343 (2025)
29. Lin, W., Wei, J., Liu, B., Zhang, Y., Yan, S., Guo, M.: Cascadev: An implementation of wurstchen architecture for video generation. arXiv preprint arXiv:2501.16612 (2025)
30. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., Liu, X., Yang, F., Wan, P., ZHANG, D., Gai, K., Yang, Y., Ouyang, W.: Improving video generation with human feedback. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=nHkg4yc7SP

31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
32. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
33. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14297–14306 (2023)
34. Mi, X., Yu, W., Lian, J., Jie, S., Zhong, R., Liu, Z., Zhang, G., Zhou, Z., Xu, Z., Zhou, Y., Lu, Q., Tang, F.: Video generation models are good latent reward models. arXiv preprint arXiv:2511.21541 (2025)
35. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=B1QRgziT-
36. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
37. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
38. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1) (Jan 2020)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=TIdIXIpzhoI
41. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast high-resolution image synthesis with latent adversarial diffusion distillation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
42. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
43. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-a-video: Text-to-video generation without text-video data. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=nJfylDvgzlq
44. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 32211–32252. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/song23a.html
45. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=PxTIG12RRHS

46. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
47. 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)
48. Wang, F.Y., Huang, Z., Bergman, A., Shen, D., Gao, P., Lingelbach, M., Sun, K., Bian, W., Song, G., Liu, Y., et al.: Phased consistency models. Advances in neural information processing systems **37**, 83951–84009 (2024)
49. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
50. Wang, Q., Shi, Y., Ou, J., Chen, R., Lin, K., Wang, J., Jiang, B., Yang, H., Zheng, M., Tao, X., Yang, F., Wan, P., Zhang, D.: Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8428–8437 (June 2025)
51. Wang, Y., Tan, Z., Wang, J., Yang, X., Jin, C., Li, H.: Lift: Leveraging human feedback for text-to-video model alignment. arXiv preprint arXiv:2412.04814 (2024)
52. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2096–2105 (2023)
53. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)
54. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
55. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8941–8951 (2024)
56. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
57. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. 2024 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6613–6623 (2023)
58. Zhang, J., Wu, J., Ren, Y., Xia, X., Kuang, H., Xie, P., Li, J., Xiao, X., Huang, W., Wen, S., et al.: Unifl: Improve latent diffusion model via unified feedback learning. Advances in Neural Information Processing Systems **37**, 67355–67382 (2024)
59. Zhang, J., Zheng, K., Jiang, K., Wang, H., Stoica, I., Gonzalez, J.E., Chen, J., Zhu, J.: Turbodiffusion: Accelerating video diffusion models by 100-200 times. arXiv preprint arXiv:2512.16093 (2025)
60. Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., et al.: Pytorch fsdp: experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277 (2023)

61. Zheng, K., Wang, Y., Ma, Q., Chen, H., Zhang, J., Balaji, Y., Chen, J., Liu, M.Y., Zhu, J., Zhang, Q.: Large scale diffusion distillation via score-regularized continuous-time consistency. arXiv preprint arXiv:2510.08431 (2025)
62. Zheng, K., Wang, Y., Ma, Q., Chen, H., Zhang, J., Balaji, Y., Chen, J., Liu, M.Y., Zhu, J., Zhang, Q.: Large scale diffusion distillation via score-regularized continuous-time consistency. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=2uNlM353RI