

Video Generation Models Are Good Latent Reward Models

Xiaoyue Mi^{1,†,*} Wenqing Yu^{2,*} Jiesong Lian³ Shibo Jie⁴ Ruizhe Zhong⁵
Zijun Liu⁶ Guozhen Zhang⁷ Zixiang Zhou² Zhiyong Xu²
Yuan Zhou^{2,‡} Qinglin Lu² Fan Tang^{1,§}

¹University of Chinese Academy of Sciences ²Tencent Hunyuan

³Huazhong University of Science and Technology

⁴Peking University ⁵Shanghai Jiao Tong University ⁶Tsinghua University ⁷Nanjing University

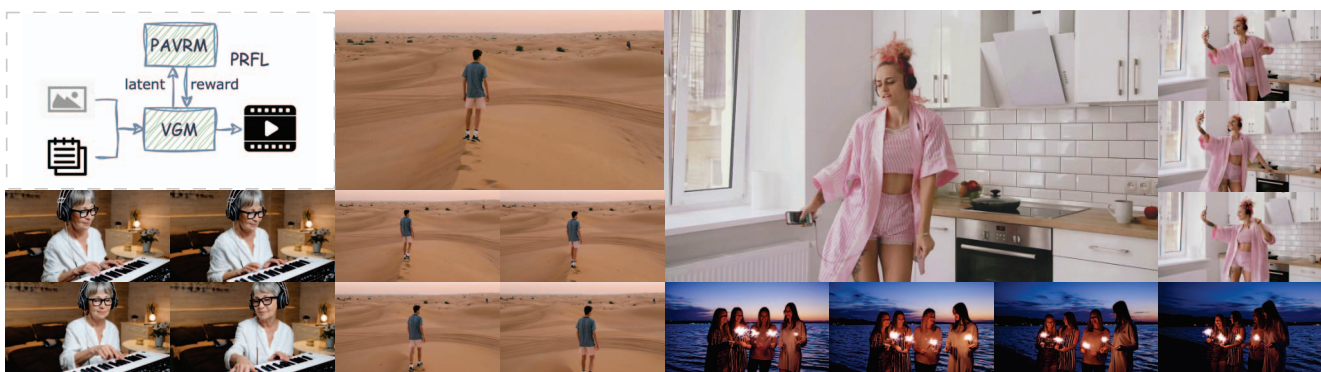


Figure 1. PRFL generates high-quality videos with enhanced motion quality. We show results on image-to-video (frames of different sizes) and text-to-video (frames of same size) generation tasks, showing the representative frames from each video.

Abstract

Reward feedback learning (ReFL) has proven effective for aligning image generation with human preferences. However, its extension to video generation faces significant challenges. Existing video reward models rely on vision-language models designed for pixel-space inputs, confining ReFL optimization to near-complete denoising steps after computationally expensive VAE decoding. This pixel-space approach incurs substantial memory overhead and increased training time, and its late-stage optimization lacks early-stage supervision, refining only visual quality rather than fundamental motion dynamics and structural coherence. In this work, we show that pre-trained video generation models are naturally suited for reward modeling in the noisy latent space, as they are explicitly designed to process noisy latent representations at arbitrary timesteps and inherently preserve temporal information through their sequential modeling capabilities. Ac-

cordingly, we propose *Process Reward Feedback Learning (PRFL)*, a framework that conducts preference optimization entirely in latent space, enabling efficient gradient backpropagation throughout the full denoising chain without VAE decoding. Extensive experiments demonstrate that PRFL significantly improves alignment with human preferences, while achieving substantial reductions in memory consumption and training time compared to RGB ReFL. Code is available at <https://github.com/Tencent-Hunyuan/HY-Video-PRFL>

1. Introduction

Recent video generation models [5, 12, 14, 25] have demonstrated remarkable visual fidelity, producing photorealistic content across diverse applications [9, 18, 22]. Nevertheless, aligning these models with complex human preferences, particularly concerning motion quality [17], physical plausibility [1], and prompt following [2], remains a fundamental challenge [32]. Reward feedback learning (ReFL) [3, 23, 30] has emerged as a promising paradigm for preference alignment in image generation, yet its direct application to video generation encounters critical compu-

[†] Work done during internship at Tencent Hunyuan.

^{*} Equal contribution. mxysdu@gmail.com

[‡] Project leader.

[§] Corresponding author. tfan.108@gmail.com

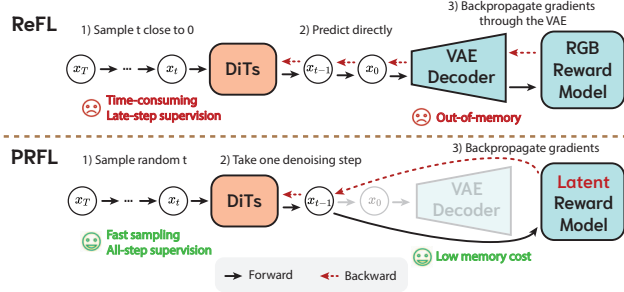


Figure 2. Comparison between RGB ReFL and our PRFL. RGB reward models require near-complete denoising and VAE decoding to RGB space, introducing *evaluation delay*, *GPU memory bottleneck*, and *insufficient supervision* of early denoising stages where structure and motion are formed. PRFL eliminates these limitations by performing reward modeling directly in latent space with timestep-aware training.

tational and optimization barriers that fundamentally limit its applicability.

Typical ReFL approaches rely on outcome-based reward models built upon Vision-Language Models (VLMs) [7, 8, 17, 27, 30]. As shown in Fig. 2, such outcome-based reward models operate on videos in RGB space, imposing three interconnected limitations due to their inherent post-hoc nature. **Initially**, near-complete denoising required by VLMs introduces significant *evaluation delay*. **Furthermore**, backpropagation through the VAE decoder causes a *GPU memory bottleneck*. **Ultimately**, rewards applied only at the final step yield *outcome-level supervision*, failing to guide early stages where motion structure is formed. Although recent methods [6, 28] attempt partial relief via gradient stopping or trajectory shortcuts, they still depend on fully denoised RGB frames. ContentV [15] bypasses VAE decoding by optimizing only the first frame yet sacrifices holistic video quality assessment, while DOLLAR [4] employs a VLM-based latent reward model but lacks fine-grained timestep-wise supervision. These limitations share a common root: all existing approaches remain structurally anchored to outcome-level reward, leaving the denoising process itself without any direct supervision.

A natural question arises: *can the video generation model itself serve as a quality-aware critic, providing reward signals directly within the denoising process?* As shown in Fig. 3, linear probing on VGM latent features reveals strong discriminability of human-perceived video quality at arbitrary denoising timesteps, even at highly noisy early stages where no visual content is yet interpretable, suggesting that VGMs inherently encode quality-relevant information throughout the denoising trajectory. Yet translating this observation into a practical reward model raises two non-trivial challenges: how to aggregate variable-length spatiotemporal features into a quality-aware representation while maintaining noise-level sensitivity, and how

to distribute process rewards across full denoising trajectories without rollout overhead. The closest prior works fall short on both fronts: LPO [33] operates on image latents and cannot generalize to variable-length spatiotemporal video tensors, while VideoAlign [17] targets inference-time guidance and lacks the training-time dense supervision that both challenges demand. To our knowledge, training-time process reward supervision over full video denoising trajectories remains an open problem.

In this work, we address both challenges through process-level reward modeling and optimization. We propose **Process-Aware Video Reward Model (PAVRM)**, which repurposes pretrained VGMs to evaluate quality directly from noisy latents at arbitrary timesteps. PAVRM employs learnable query vectors to compress variable-length spatiotemporal features into a compact quality-aware token, inherently encouraging the model to capture quality-relevant patterns rather than memorizing content-specific correlations. We further introduce **Process Reward Feedback Learning (PRFL)**, which leverages PAVRM as a training-time critic by sampling random timesteps and maximizing process rewards through a single denoising gradient, entirely eliminating VAE decoding overhead while distributing supervision across the full denoising trajectory. Extensive experiments demonstrate that PRFL achieves **substantial improvements** (up to **+56.00** in dynamic degree, **+21.52** in human anatomy) with **significant memory savings** and at least **1.4× faster training** over RGB-space ReFL baselines.

Our contributions can be summarized as follows:

- We propose a **process-aware video reward model (PAVRM)**, which employs query-based aggregation to efficiently handle variable-length videos with timestep sensitivity, and keep artifact awareness throughout the denoising process based on pre-trained video models.
- We introduce **PRFL**, the first training-time post-training framework that provides process-level supervision across the full video denoising trajectory, eliminating VAE decoding overhead through single-step latent optimization at randomly sampled timesteps.
- Experiments show our approach improves motion quality while saving substantial GPU memory and at least accelerating training by 1.4× compared to RGB ReFL.

2. Related Work

2.1. Visual Reward Feedback Learning

Reward feedback learning has proven effective for improving visual generation quality. Early works [3, 23, 30] fine-tune diffusion models using differentiable reward signals, but face a depth-efficiency dilemma: full backpropagation incurs prohibitive memory costs, while training only final steps fails to optimize low-level objectives. Recent ap-

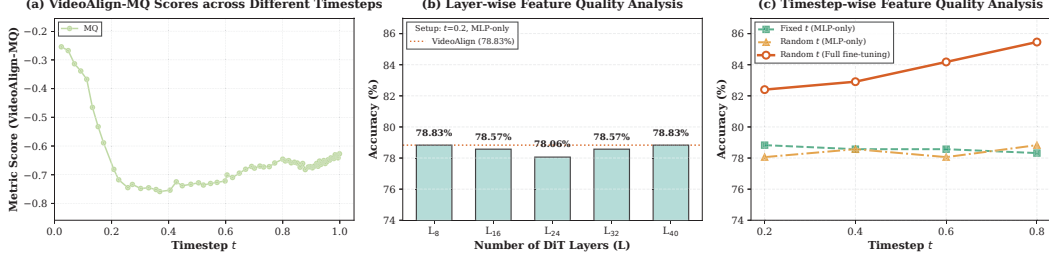


Figure 3. Analysis on VGM Features. (a) VLM-based reward model (VideoAlign-MQ) exhibits poor timestep generalization with fluctuating scores. (b) VGM features from any DiT layer uniformly achieve 78.8% accuracy, matching VLM baseline. (c) Timestep-aware fine-tuning unlocks VGM’s full potential, achieving 85.46% accuracy with peak performance at early timesteps ($t=0.8$).

proaches [6, 28] mitigate this via gradient stopping or trajectory shortcuts, yet remain anchored to pixel-space rewards that require VAE decoding, a cost amplified severely in video generation due to multi-frame synthesis. ContentV [15] avoids VAE decoding by optimizing only the first frame but sacrifices holistic video quality, while DOLLAR [4] employs a VLM-based latent reward for distillation yet lacks timestep-wise optimization. We propose a latent reward model enabling gradient backpropagation at arbitrary timesteps without VAE decoding.

2.2. Reward Models for Video Generation

Reward models are categorized into outcome models [24] supervising final results and process models [13, 26, 34] supervising intermediate steps; current video reward models focus exclusively on the former. Early video reward models adapted image-based approaches [31] or introduced VLMs for quality assessment [7, 8, 17, 21, 29], but operating only on near-complete outputs leaves early denoising stages, where fundamental video attributes are established, without direct supervision. LPO [33] pioneered diffusion models as noise-aware latent reward models for images, yet cannot generalize to variable-length spatiotemporal video tensors. VideoAlign [17] briefly explored VGMs for inference-time guidance, but training-time process reward supervision remains unexplored. We systematically investigate VGMs as process reward models, offering timestep-aware quality evaluation with native support for full-sequence video processing.

3. Preliminaries and Feasibility Analysis

Rectified Flow. Recently, video generation models [5, 12, 25] operate in rectified flow [16, 19], which establishes a continuous transport between data and noise distributions. Specifically, given clean data $\mathbf{x}_0$ sampled from distribution $q(\mathbf{x}_0)$ and noise $\mathbf{x}_1$ from distribution $p(\mathbf{x}_1)$, the framework defines a time-dependent interpolation path:

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1, \quad t \in [0, 1]. \quad (1)$$

A neural network (usually several DiT blocks) parameterized by θ predicts the velocity field $\mathbf{v}_\theta(\mathbf{x}_t, t)$ at each

timestep, trained via the flow matching objective, which also as supervised fine-tuning (SFT) loss:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,1), \mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{x}_1 \sim p(\mathbf{x}_1)} [\|\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{v}\|^2], \quad (2)$$

where $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$ denotes the data-to-noise transport direction.

Reward Feedback Learning. Reward Feedback Learning (ReFL) [30] optimizes diffusion models by backpropagating reward signals through the denoising process. Given a reward model r_ϕ and generated output $\mathbf{x}_0$, ReFL samples a timestep t from the late denoising stage and computes gradients from the reward score. To prevent reward over-optimization, the objective combines reward-based loss with SFT regularization:

$$\mathcal{L}_{\text{ReFL}} = -\lambda \mathbb{E}_{\mathbf{x}_0 \sim \text{VGM}_\theta} [r_\phi(\mathcal{D}(\mathbf{x}_0))] + \mathcal{L}_{\text{FM}}(\theta), \quad (3)$$

where $\mathcal{D}$ denotes the VAE decoder.

Feasibility Analysis of VGM-based Reward Models.

We analyze whether VGMs can serve as effective reward models using motion quality as a case study.

VLM fails on noisy inputs (Fig. 3(a)). VideoAlign [17], a representative VLM-based video reward model trained on large-scale preference data, exhibits severe performance degradation across different denoising timesteps. When evaluating latents decoded to RGB space, the logit scores fluctuate dramatically between early and late timesteps, revealing poor generalization to high-noise regions.

VGM features encode motion efficiently (Fig. 3(b)). We employ a simple MLP as a linear probe to assess VGM feature quality. Remarkably, probing features from any DiT layer (L_8 to L_{40}) achieves uniform performance at 78.8%, matching the VideoAlign baseline. This confirms that motion dynamics are efficiently distributed throughout the network, enabling the use of only the first 8 layers for a more efficient reward model.

Timestep-aware fine-tuning unlocks VGM’s potential (Fig. 3(c)). While MLP-only probing (both fixed and random timestep) fails to surpass the VLM baseline, coupling full fine-tuning with random timestep sampling yields

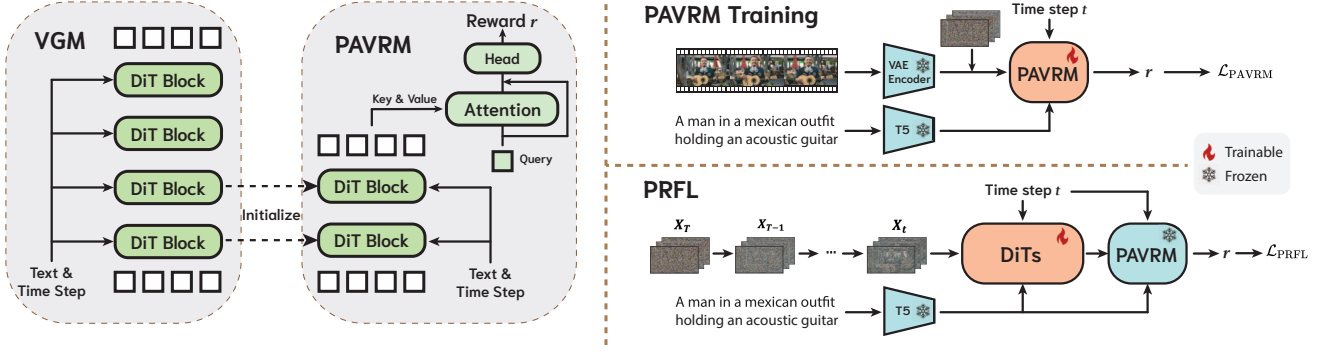


Figure 4. Overview of our process-aware video generation alignment framework. Left: Architecture of the Video Generation Model (VGM) and Process-Aware Video Reward Model (PAVRM). Right: Two-stage training pipeline. PAVRM training with reward prediction from noisy latents. Process Reward Feedback Learning (PRFL) optimizing VGM through latent-space reinforcement learning at randomly sampled timesteps.

Algorithm 1 Process Reward Feedback Learning (PRFL)

```

1: Input: Velocity predictor  $\mathbf{v}_\theta$ , PAVRM  $F_\phi$ , VAE encoder  $\mathcal{E}$ ,
   datasets  $\mathcal{D}_{\text{PRFL}}$ ,  $\mathcal{D}_{\text{SFT}}$ , reward weight  $\lambda$ , total timesteps  $T$ ,
   timestep interval  $\Delta t$ , learning rate  $\eta$ 
2: for each training iteration do
3:   // Supervised Fine-Tuning Regularization
4:   Sample  $(V, I, p) \sim \mathcal{D}_{\text{SFT}}$  and  $t_{\text{sft}} \sim \mathcal{U}(0, 1)$ 
5:   Compute  $\mathbf{x}_0 = \mathcal{E}(V)$ , sample  $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$ 
6:   Construct  $\mathbf{x}_{t_{\text{sft}}} = (1 - t_{\text{sft}})\mathbf{x}_0 + t_{\text{sft}}\mathbf{x}_1$ 
7:   Compute  $\mathcal{L}_{\text{FM}} = \|\mathbf{v}_\theta(\mathbf{x}_{t_{\text{sft}}}, t_{\text{sft}}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2$ 
8:   Update  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{FM}}$ 
9:   // Process Reward Feedback Learning
10:  Sample  $(I, p) \sim \mathcal{D}_{\text{PRFL}}$ ,  $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ , and  $t \sim \mathcal{U}(0, T - 1)$ 
11:  for  $j = T, T - 1, \dots, t + 1$  do
12:    no grad:  $\mathbf{x}_{j-1} \leftarrow \mathbf{x}_j - \Delta t \cdot \mathbf{v}_\theta(\mathbf{x}_j, j)$ 
13:  end for
14:  with grad:  $\mathbf{x}_t \leftarrow \mathbf{x}_{t+1} - \Delta t \cdot \mathbf{v}_\theta(\mathbf{x}_{t+1}, t + 1)$ 
15:  Compute  $r_t = F_\phi(\mathbf{x}_t, t, I, p)$  and  $\mathcal{L}_{\text{PRFL}} = -\lambda \cdot r_t$ 
16:  Update  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{PRFL}}$ 
17: end for

```

a dramatic leap to **85.46%** ($\sim 6.6\%$ absolute gain), particularly when sampling features from earlier noisy states ($t=0.8$). This validates that fine-tuned VGMs provide superior, timestep-robust signals for learning motion quality rewards.

4. Method

4.1. Overview

We propose process reward feedback learning, a new ReFL framework in the video generation task to address GPU memory and training time limitations. As shown in Fig. 4, the full process consists of two stages: (1) training a Process-Aware Video Reward Model (PAVRM, Sec. 4.2) adapted from the VGM that evaluates video quality directly from noisy latents at arbitrary diffusion timesteps, and (2) optimizing the VGM using Process Reward Feed-

back Learning (PRFL, Sec. 4.3).

4.2. Process-Aware Video Reward Model

Architecture of PAVRM. PAVRM leverages features from a frozen pre-trained VGM to predict preference scores directly from noisy latents, as illustrated in Fig. 4 (left). Given a noisy latent video $\mathbf{x}_t \in \mathbb{R}^{F \times H \times W \times C}$ at timestep t and a text prompt p , we use the first eight DiT blocks of the backbone as **spatiotemporal feature-extractor**:

$$\mathbf{h} = \text{DiT}_\phi(\mathbf{x}_t, t, \mathcal{T}(p)) \in \mathbb{R}^{F \times H \times W \times D}, \quad (4)$$

where $\mathcal{T}(\cdot)$ denotes the text encoder, and D denotes the feature dimension. These intermediate representations naturally encode motion quality signals at various noise levels, as demonstrated in Fig. 3.

To handle variable-length video features and extract discriminative representations, we employ a query attention as **query-based spatiotemporal aggregation** that adaptively compresses spatiotemporal information into a fixed-size embedding. Specifically, we flatten the spatiotemporal features $\mathbf{h} \in \mathbb{R}^{F \times H \times W \times D}$ into $\hat{\mathbf{h}} \in \mathbb{R}^{N \times D}$ where $N = F \cdot H \cdot W$, and compute:

$$\mathbf{z}_{\text{obs}} = \frac{\exp(\mathbf{q}(\hat{\mathbf{h}}\mathbf{W}_K)^T / \sqrt{D})}{\sum \exp(\mathbf{q}(\hat{\mathbf{h}}\mathbf{W}_K)^T / \sqrt{D})} (\hat{\mathbf{h}}\mathbf{W}_V) \in \mathbb{R}^{1 \times D}, \quad (5)$$

where $\mathbf{q} \in \mathbb{R}^{1 \times D}$ is a learnable query vector, and $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times D}$ are projection matrices. The final representation $\mathbf{z} = \mathbf{z}_{\text{obs}} + \mathbf{q} \in \mathbb{R}^D$ concatenates observation features with content-agnostic quality priors, enabling the model to reason about generation quality independent of content correlations. A three-layer MLP as **head layer** maps $\mathbf{z}$ to the reward score $r_\phi(\mathbf{x}_t, t, p) \in \mathbb{R}$.

Training Objective. PAVRM is trained on a binary preference dataset $\mathcal{D}_{\text{RM}} = \{(V_i, p_i, y_i)\}_{i=1}^N$ where $y_i \in \{0, 1\}$ indicates motion quality satisfaction. Following the rectified flow formulation (Eq. (1)), we construct noisy latents

$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1$ with $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$ and $t \sim \mathcal{U}(0, 1)$, where $\mathbf{x}_0 = \mathcal{E}(V)$. The training objective minimizes:

$$\mathcal{L}_{\text{PAVRM}} = -\mathbb{E}_{t, (V, p, y)} \left[y \log \sigma(r_\phi(\mathbf{x}_t, t, p)) + (1 - y) \log(1 - \sigma(r_\phi(\mathbf{x}_t, t, p))) \right], \quad (6)$$

where $t \sim \mathcal{U}(0, 1)$ and $(V, p, y) \sim \mathcal{D}_{\text{RM}}$. We freeze the VAE encoder $\mathcal{E}$ while jointly optimizing the DiT blocks, query vector $\mathbf{q}$, and MLP parameters.

A critical design of PAVRM is the random sampling of timesteps $t \in [0, 1]$ during training. Unlike outcome-based reward models that only evaluate clean outputs ($t \approx 0$), PAVRM learns stage-appropriate quality assessment across the entire denoising trajectory. This is justified in rectified flow: since $\mathbf{x}_t$ follows a deterministic linear interpolation from noise to data, preference labels of final outputs naturally propagate to intermediate states. This process-aware training enables effective guidance throughout the generation process, as demonstrated in Sec. 5.

4.3. Process Reward Feedback Learning

As discussed in Sec. 3, vanilla RGB ReFL (Eq. (3)) optimizes only final denoised outputs $\mathbf{x}_0$, requiring VAE decoding $\mathcal{D}(\mathbf{x}_0)$ and pixel-space reward computation—prohibitively expensive for high-resolution videos. Moreover, sampling timesteps only from late denoising stages fails to improve early-stage motion planning, which critically determines overall physical coherence. PRFL addresses these limitations by performing reward optimization at arbitrary intermediate timesteps directly in latent space, eliminating VAE decoding while distributing learning signals across the entire generation trajectory.

Training Procedure. As shown in Alg. 1, we extend the ReFL framework by replacing outcome rewards with process-level rewards from PAVRM. Given the text prompt p from the dataset $\mathcal{D}_{\text{PRFL}}$, we sample the initial noise $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$ and the target timestep $s \sim \mathcal{U}(0, 1)$. In practice, we perform gradient-free denoising rollouts from $t = 1$ to $t = s + \Delta t$, and then execute one gradient-enabled step:

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t - \Delta t \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, p) (\text{w/o. grad}), \quad (7)$$

$$\mathbf{x}_s = \mathbf{x}_{s+\Delta t} - \Delta t \cdot \mathbf{v}_\theta(\mathbf{x}_{s+\Delta t}, s + \Delta t, p) (\text{w/ grad}), \quad (8)$$

where $\Delta t = \frac{1}{N}$ denotes the discrete step size with N total denoising steps. The process reward loss directly maximizes PAVRM-predicted quality at intermediate timestep s :

$$\mathcal{L}_{\text{PRFL}} = -\lambda \mathbb{E}_{s, p} [r_\phi(\mathbf{x}_s, s, p)], \quad (9)$$

where λ controls optimization strength. Crucially, gradients $\nabla_\theta \mathcal{L}_{\text{PRFL}}$ backpropagate through Eq. (8) and PAVRM r_ϕ

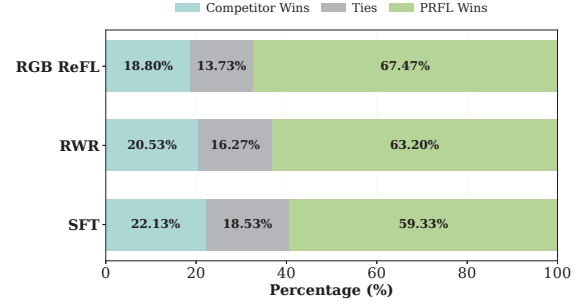


Figure 5. Human evaluation of PRFL model vs. other post-training methods.

without requiring VAE decoding $\mathcal{D}(\cdot)$, enabling memory-efficient training.

Following the regularization strategy in Eq. (3), we prevent reward over-optimization by alternating with supervised fine-tuning on curated dataset $\mathcal{D}_{\text{SFT}} = \{(V_i, p_i)\}_{i=1}^M$:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{t, (V, p)} \|\mathbf{v}_\theta(\mathbf{x}_t, t, p) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2, \quad (10)$$

where $\mathbf{x}_0 = \mathcal{E}(V)$, $t \sim \mathcal{U}(0, 1)$, and $(V, p) \sim \mathcal{D}_{\text{SFT}}$. This balanced training strategy, alternating between $\mathcal{L}_{\text{PRFL}}$ and $\mathcal{L}_{\text{SFT}}$ at each iteration, maintains generation diversity while adapting to motion quality preferences.

5. Experiments

5.1. Experimental Setup

Dataset. We collected about 31,000 portrait videos from online sources and generated text prompts using a video caption model. The first frame and text prompt of each video are fed into the Wan2.1-14B-I2V model for inference. The generated videos are labeled by professional annotators based on motion quality as qualified, partially qualified, and unqualified. The partially qualified videos are filtered out to enhance distinctiveness. Detailed annotation and filtration guidelines are provided in the supplementary materials. This process yielded a dataset comprising 24,000 real videos and corresponding generated videos. For reward model, we utilized only the generated videos, and randomly selected 500 samples as the test set (100 for validation, 400 for testing). For video generation, we employed only real videos as SFT training data, and we randomly selected 100 input conditions as our test dataset from the 500-sample test set of the reward model. We also incorporated the existing open-source benchmark VBench [10] and VBench2 [36] to ensure a fair comparison with state-of-the-art methods.

Training Details. We adopt Wan2.1-I2V-14B [25] as our primary baseline. We employ the AdamW optimizer [20] with learning rates of 1e-5 for query attention and head of PAVRM, 1e-6 for the feature extraction of PAVRM, and 5e-6 for PRFL. We utilize UniPCMultistepScheduler [35] with 1,000 training steps and 40 inference steps.

Table 1. **Comprehensive comparison across video generation benchmarks on text-to-video generation task.** Evaluation is conducted on Inner Test Set, VBench and VBench2. For Inner Test Set, we evaluate MS (motion smoothness), DD (dynamic degree), SC (subject consistency) from VBench, HA (human anatomy) from VBench2, and our proposed PAVRM. Base model is Wan2.1-T2V-14B. **Blue** denotes the best results. **Blue** values indicate absolute improvements of PRFL over Pretrain baseline. PRFL achieves significant improvements in dynamic degree and human anatomy.

Method	Resolution	Inner Test Set					VBench				VBench2		Avg
		MS	DD	SC	HA	PAVRM	MS	DD	SC	PAVRM	HA	PAVRM	
Pretrain [25]	480P	99.20	22.00	97.34	84.24	89.00	98.00	68.06	92.74	97.22	74.38	69.17	81.03
SFT	480P	98.96	44.00	96.61	92.79	92.00	97.87	62.50	93.21	95.83	84.80	74.17	84.79
RWR [17]	480P	98.99	60.00	95.93	91.85	88.00	98.06	65.28	92.48	93.06	79.67	62.50	84.17
RGB ReFL [15]	480P	99.20	38.00	92.26	91.68	92.00	98.64	62.50	90.60	97.22	84.87	76.67	83.97
PRFL	480P	99.05	68.00	96.34	94.73	92.00	98.18	76.39	94.16	100.00	89.84	76.67	89.58
vs. Pretrain		-0.15	+46.00	-1.00	+10.49	+3.00	+0.18	+8.33	+1.42	+2.78	+15.46	+7.50	+8.55
Pretrain [25]	720P	99.09	25.00	96.69	78.73	94.00	97.70	61.11	90.63	98.61	68.88	62.10	79.32
PRFL	720P	98.85	81.00	96.09	90.89	95.00	98.06	84.72	95.46	100.00	90.40	66.13	90.60
vs. Pretrain		-0.24	+56.00	-0.60	+12.16	+1.00	+0.36	+23.61	+4.83	+1.39	+21.52	+4.03	+11.28

Table 2. **Comprehensive comparison across video generation benchmarks on image-to-video generation task.** Evaluation is conducted on Inner Test Set and VBench. For Inner Test Set, we evaluate MS (motion smoothness), DD (dynamic degree), SC (subject consistency), IC (i2v subject) from VBench-I2V, and our proposed PAVRM. **Blue** denotes the best results. **Blue** values indicate absolute improvements of PRFL over Pretrain baseline. PRFL achieves significant improvements in dynamic degree.

Method	Backbone	Inner Test Set					VBench-I2V					Avg
		MS	DD	SC	IC	PAVRM	MS	DD	SC	IC	PAVRM	
Pretrain [25]	Wan2.1-I2V-14B-480P	98.66	57.00	91.73	96.86	87.00	97.86	40.65	93.86	97.21	92.28	85.31
PRFL	Wan2.1-I2V-14B-480P	98.88	87.00	93.18	97.31	93.00	98.04	81.30	94.57	97.79	92.68	93.38
vs. Pretrain		+0.22	+30.00	+1.45	+0.45	+6.00	+0.18	+40.65	+0.71	+0.58	+0.40	+8.07
Pretrain [25]	Wan2.1-I2V-14B-720P	98.40	60.00	90.96	96.65	74.00	98.04	35.37	94.49	97.92	89.43	83.53
PRFL	Wan2.1-I2V-14B-720P	99.03	76.00	92.83	98.26	90.00	98.65	68.42	95.62	98.73	95.53	91.31
vs. Pretrain		+0.63	+16.00	+1.87	+1.61	+16.00	+0.61	+33.05	+1.13	+0.81	+6.10	+7.78

Comparison. For reward model, we select VideoAlign [17] and VideoPhy [1], a state-of-the-art RGB-based reward model. For video generation, we compare against four post-training methods, two offline methods: SFT [19] and reward weighted regression (RWR) [17], and one online RL method RGB ReFL [15], which only decodes the first frame and uses PickScore [11] as reward model. See details in supplementary materials.

Evaluation. For PAVRM evaluation, we conduct stratified random sampling of timestep t across five intervals: $[0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, 1.0]$. Each test sample is randomly sampled once within each interval, and the final average accuracy is computed. For video generation evaluation, generated videos are produced at 720P/480P resolution. For the text-to-video generation task, we employ evaluation metrics from the VBench series, including dynamic degree, motion smoothness, subject consistency, and Human Anatomy, as well as the qualified ratio across the whole test set estimated by PAVRM. Specifically, for each test sample, we random sample a timestep t across $[0, 1.0]$ fed into PAVRM and get predictions 0 as unqualified or 1 as qualified. For the image-to-video generation task, additional i2v subject metric is used from the VBench i2v series.

5.2. Video Generation

Quantitative Results. As shown in Tables 1 and 2, we evaluate PRFL across different resolutions and tasks.

Text-to-Video Generation. In the T2V 480P setting, PRFL substantially outperforms other post-training methods (SFT, RWR, RGB ReFL) in dynamic degree (+46.00 on Inner Test Set) and human anatomy (+10.49), while maintaining high performance in motion smoothness (99.05) and subject consistency (96.34). The slight decreases in MS (-0.15) and SC (-1.00) are marginal given their already high baseline values (99.20 and 97.34). This suggests PRFL enhances motion dynamics without significantly compromising smoothness or consistency. Similar trends are observed at 720P resolution, with even larger improvements in dynamic degree (+56.00) and human anatomy (+12.16).

Image-to-Video Generation. PRFL generalizes well to the I2V task. At 480P, dynamic degree improves by +30.00 on Inner Test Set and +40.65 on VBench-I2V, while other metrics either improve or remain stable (MS: +0.22, SC: +1.45, IC: +0.45). The 720P results confirm consistent improvements across metrics. These results indicate that PRFL’s effectiveness extends beyond T2V generation.

The consistent improvements in motion-related metrics (dynamic degree, human anatomy) across different settings, combined with the preservation of high-level smoothness

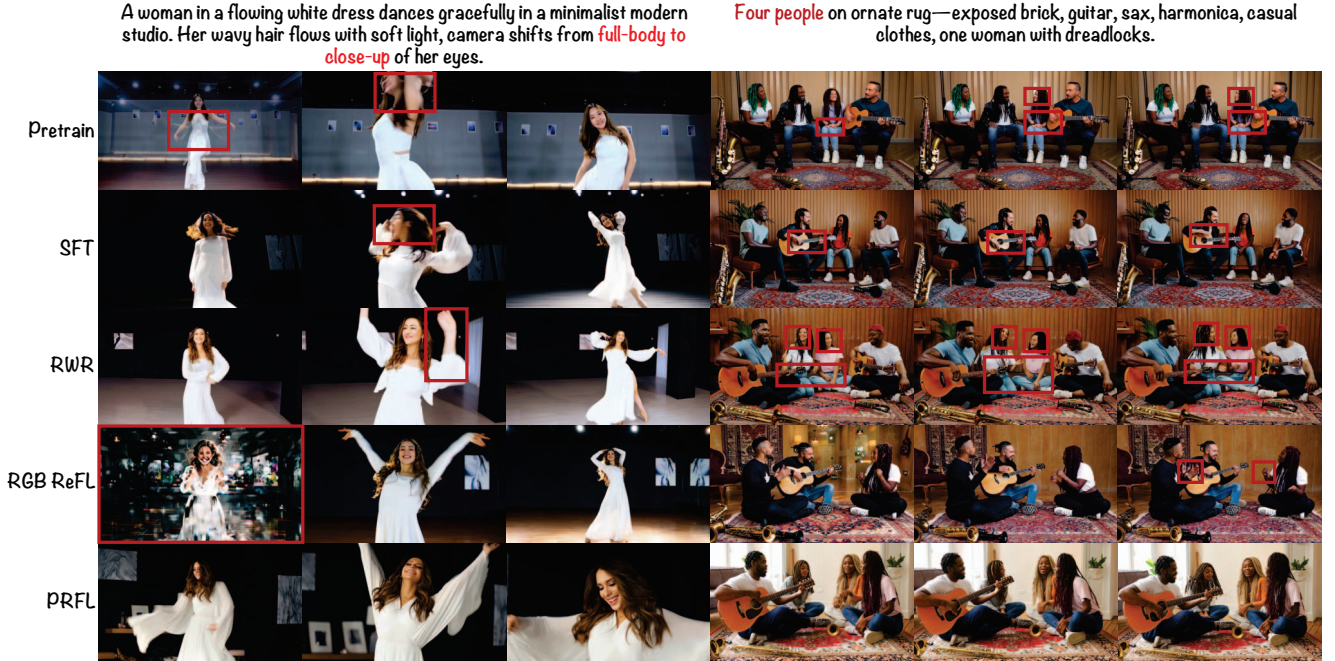


Figure 6. Qualitative results for different post-training methods on 480P text-to-video task. The red box highlights the generated artifacts. Zoom in for a better view. For the complete prompt, please see the supplementary materials.

Table 3. **Analysis of the influence of different timestep-sampling strategies.** Base model is Wan2.1-T2V-14B at 480P resolution. The early, middle, and late stages of the denoising process refer to the first third, middle third, and final third of the process, respectively.

Method	Inner Test Set					VBench				VBench2		Avg
	MS	DD	SC	HA	PAVRM	MS	DD	SC	PAVRM	HA	PAVRM	
Pretrain	99.20	22.00	97.34	84.24	89.00	98.00	68.06	92.74	97.22	74.38	69.17	81.03
Early Stage	99.00	51.00	96.63	87.52	85.00	98.53	48.61	96.71	95.83	78.47	67.50	82.25
Middle Stage	99.11	51.00	96.69	89.38	92.00	98.32	76.39	94.84	98.61	80.92	80.00	87.02
Late Stage	99.26	44.00	96.29	91.54	93.00	98.60	58.33	93.94	98.61	84.07	77.50	85.01
Full Stage	99.05	68.00	96.34	94.73	92.00	98.18	76.39	94.16	100.00	89.84	76.67	89.58

Table 4. Comparison of training resource consumption and efficiency. PRFL achieves $1.4\times$ speedup while processing full frames. Here † means without SFT loss, and **bold** means the best performance.

Method	VRAM (GB)	Time per Step (s)	Speedup
RGB ReFL (full frames)	OOM	-	-
RGB ReFL (first frame)	55.47	72.38 / 64.89 †	1.00 $\times$
PRFL (full frames)	66.81	51.11 / 43.69†	1.42$\times$ / 1.49$\times$

and consistency, demonstrate the effectiveness of PRFL. While automatic metrics have limitations in capturing perceptual quality, the quantitative results provide evidence that PRFL successfully optimizes motion quality across diverse video generation scenarios.

User Study. To complement automatic metrics, we conduct a comprehensive human evaluation study. We randomly sample 25 prompts from the test set and generate videos using PRFL and three baseline methods (SFT, RWR, RGB ReFL). We recruit 2,250 pairwise comparisons based on overall video quality from 30 professional partic-

ipants. Each participant evaluates video pairs and chooses A wins/tie/B wins based on comprehensive consideration of text instruction adherence, visual quality, and motion quality. As shown in Fig. 5, PRFL consistently outperforms all baseline methods. These results demonstrate that PRFL’s process-level optimization effectively improves video quality as perceived by human evaluators.

Qualitative Results. As shown in Fig. 6, we visualize three representative frames (early, middle, and late stages) from generated videos to compare PRFL with baseline methods. In the dancing scenario with complex motion dynamics, baseline methods exhibit various artifacts: Pretrain generates a distorted environment and figure, SFT produces facial distortions in the close-up frame, RWR shows body deformations, and RGB ReFL generates a failed first frame. In contrast, PRFL generates consistent and artifact-free frames throughout the video while maintaining smooth motion transitions and accurately following the text prompt’s camera movement requirements. In the multi-person gui-

Table 5. Quantitative results for different architecture PAVRMs across different timesteps on T2V and I2V tasks under 480P and 720P. The metric is average accuracy calculated after randomly sampling a time step within each time step range for every test sample. **Bold** means best performance.

Method	Base Model	Task	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]	Avg
<i>Resolution: 720P</i>								
VideoAlign [17]	VLM	I2V	-	-	-	-	-	78.83
VideoPhy [1]	VLM	I2V	-	-	-	-	-	77.04
Mean Pooling	VGM	I2V	82.40	82.91	83.67	85.46	82.40	83.37
Max Pooling	VGM	I2V	80.61	80.10	80.61	77.30	73.72	78.47
Attention w/o. query	VGM	I2V	80.36	84.44	84.95	84.18	83.16	83.42
Attention w./ query	VGM	I2V	83.42	84.95	84.69	84.44	83.42	84.18
Attention w./ query	VGM	T2V	82.40	84.69	84.69	84.44	84.44	84.13
<i>Resolution: 480P</i>								
Attention w./ query	VGM	I2V	83.67	84.18	84.69	84.69	79.85	83.42
Attention w./ query	VGM	T2V	81.89	83.93	84.18	83.93	83.16	83.42

tar scene, baseline methods struggle with anatomical correctness (e.g., distorted hands and faces highlighted in red boxes), whereas PRFL maintains visual quality and generates anatomically plausible human figures and objects across all frames. These qualitative comparisons demonstrate that PRFL effectively reduces generation artifacts while preserving motion coherence in challenging scenarios. We also provide many human and non-human video cases in the supplementary materials.

Sensitivity to Sampling Timesteps. We analyze how different denoising stages affect generation quality by training PRFL with timestep sampling restricted to early, middle, and late stages of the denoising process, comparing against full-range sampling. As shown in Table 3, early and middle stages primarily govern dynamic degree and motion quality—both achieve DD of 51.00 on Inner Test Set, with the middle stage showing stronger VBench improvements (76.39 vs. 48.61). Late-stage sampling contributes less to dynamics (DD: 44.00) but substantially improves human anatomy (HA: 84.24→91.54). Full-range sampling achieves optimal overall performance (Avg: 89.58, DD: 68.00, HA: 94.73), demonstrating that optimizing across all denoising stages is essential for balancing motion and structure quality.

Computational Efficiency. As shown in Table 4, we measure the peak VRAM consumption and average time per training step for different ReFL algorithms, both with and without the SFT loss component. The results demonstrate that PRFL can decode all 81 frames while maintaining peak VRAM usage within 67 GB, which is practical for most mainstream GPU clusters. Notably, RGB-ReFL encounters out-of-memory errors when attempting to process full 81-frame sequences and must resort to first-frame-only training. Despite processing significantly more visual information (81 frames vs. 1 frame), PRFL achieves a 1.42× to 1.49× training speedup compared to the first-frame-only

RGB-ReFL baseline.

5.3. Reward Model

Architecture Selection and Timestep Analysis. As shown in Table 5, we evaluate the accuracy of PAVRM across different denoising stages and compare various aggregation methods. Attention w./ query achieves the best average performance (84.18% for I2V at 720P) and maintains stable accuracy across all timestep ranges, while simpler methods like Max Pooling suffer significant degradation in later stages. Our reward model training approach demonstrates strong generalization across different tasks (T2V vs. I2V), resolutions (480P vs. 720P), and timestep ranges, maintaining over 83% average accuracy in all evaluated settings.

6. Conclusion

We present an RLHF framework for aligning video generation models with human preferences through process-level reward modeling and optimization. Our Process-Aware Video Reward Model (PAVRM) evaluates motion quality directly from noisy latent representations at arbitrary timesteps using query-based aggregation for variable-length videos. Process Reward Feedback Learning (PRFL) enables memory-efficient fine-tuning by operating in latent space, eliminating VAE decoding overhead. Experiments demonstrate substantial motion quality improvements with large memory reduction and at least 1.4× training speedup versus pixel-space methods.

Limitations and Future Work. Our method focuses on motion quality and may benefit from multi-aspect evaluation covering aesthetics and semantics. Future work could explore hybrid reward frameworks, richer preference signals, and extensions to controllable generation and video editing.

7. Acknowledgements.

This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grant 62572458 and the Beijing Major Science and Technology Project under No. Z251100008125009.

References

- [1] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *Workshop on Video-Language Models @ NeurIPS 2024*, 2025. 1, 6, 8
- [2] Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. Vpo: Aligning text-to-video generation models with prompt optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15636–15645, 2025. 1
- [3] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2
- [4] Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17961–17971, 2025. 2, 3
- [5] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 1, 3
- [6] Xiefan Guo, Miaomiao Cui, Liefeng Bo, and Di Huang. Shortft: Diffusion model alignment via shortcut-based fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 678–687, 2025. 2, 3
- [7] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, 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*, pages 2105–2123, 2024. 2, 3
- [8] Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. In *arXiv preprint arXiv:2509.22799*, 2025. 2, 3
- [9] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025. 1
- [10] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5
- [11] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 6
- [12] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 3
- [13] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [14] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Lihuan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 1
- [15] Wenfeng Lin, Renjie Chen, Boyuan Liu, Shiyue Yan, Ruoyu Feng, Jiangchuan Wei, Yichen Zhang, Yimeng Zhou, Chao Feng, Jiao Ran, et al. Contentv: Efficient training of video generation models with limited compute. *arXiv preprint arXiv:2506.05343*, 2025. 2, 3, 6
- [16] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3
- [17] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 2, 3, 6, 8
- [18] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 1
- [19] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 6
- [20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [21] Zhun Mou, Bin Xia, Zhengchao Huang, Wenming Yang, and Jiaya Jia. Gradeo: Towards human-like evaluation for text-to-video generation via multi-step reasoning. In *Forty-second International Conference on Machine Learning*, 2025. 3
- [22] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-

- Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1
- [23] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2023. 1, 2
- [24] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3
- [25] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wenten Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3, 5, 6
- [26] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Mathshepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, 2024. 3
- [27] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 2
- [28] Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024. 2, 3
- [29] Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Boosting text-to-video generative model with mllms feedback. *Advances in Neural Information Processing Systems*, 37:139444–139469, 2024. 3
- [30] Jiazhen Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 1, 2, 3
- [31] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6463–6474, 2024. 3
- [32] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024. 1
- [33] Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 3
- [34] Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025. 3
- [35] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023. 5
- [36] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 5