GenRM-Flow: Generators are Process-aware Reward Models in Flow Matching

Siming Fu[1*] **Ruizhe He**[1*†] **Zheming Fu**[2*] **Zeyue Xue**[2] **Fengkai Liu**[2] **Yufei Jiao**[3]
Jie Huang[2] **Xiaoxiao Ma**[2] **Xuan Yang**[2] **Lihui Luo**[4]

[1]Zhejiang University [2]Joy Future Academy [3]Xiamen University [4]Tsinghua University
fusiming@zju.edu.cn 22432120@zju.edu.cn



Figure 1: Videos generated by Wan2.1-T2V-14B fine-tuned with FlowAWR using the GenRM-Flow reward, exhibiting high motion quality and dynamic fidelity.

Abstract

Aligning video generation models with human preference depends critically on the quality of the reward model that scores generated videos. Current video reward models, however, source their reward signal from a model's representational rather than generative capacity: they either repurpose vision-language models as pixel-space scorers, or train auxiliary scoring heads on top of frozen generator features. The generator's actual ability to predict velocities, denoise latents, and produce videos plays no role in scoring its own generations. We argue that this overlooks a structural property of offline preference learning. Any preference optimization objective that approximates an offline preference dataset mathematically targets a strictly defined KL-constrained optimal policy. By isolating the reward term from this analytical solution and mapping the exact generation log-likelihood to its flow-matching surrogate, the empirical reward emerges deterministically as the relative difference in continuous-time velocity-prediction errors. We formalize this equiva-

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

lence into **GenRM-Flow**, a framework where a preference-tuned video generator inherently operates as a process-aware reward model directly in the noisy latent space. We further show theoretically that distinct fine-tuning objectives, whether contrastive like DPO and IPO or non-contrastive like RWR, are alternative optimization pathways toward the same optimal policy, each inducing a reward readout from the generator's velocity-prediction errors. Furthermore, the extracted process reward seamlessly substitutes external VLM scorers in FlowAWR. Operating purely on native latent outputs without auxiliary reward networks, this generative reward drives measurable improvements in downstream video generation.

1 Introduction

Video generation models [5, 16, 54] have reached a level of fidelity where the next bottleneck is no longer raw quality but *alignment*: making the model produce videos that match human preference along axes such as motion smoothness, prompt alignment, and aesthetic quality. As in language models, alignment in video generation is increasingly framed as a reward modeling problem. Preference data is collected, a reward model (RM) is trained to score generated videos, and the generator is updated through preference optimization [28, 53] or reward feedback learning [27, 36]. The reward model is the load-bearing component of this pipeline: any error in how it scores videos translates directly into degradation of the trained generator.

Despite this central role, the shape of existing video reward models follows from a single root commitment: *the reward is computed by a model other than the generator itself*, and the rest of the architecture is largely determined by this choice. The VLM-based paradigm [28], which sources rewards from a model trained on RGB inputs, requires a noisy latent to be decoded into pixels. VAE decoding fails in high-noise regimes, restricting VLM rewards to late-stage timesteps and leaving early structure and motion formation unsupervised. The feature-extractor paradigm [36] bypasses pixels but evaluates uncalibrated intermediate features, necessitating a scoring head with independent objectives. This places the reward in a disjoint parameter space, where parametric divergence directly degrades alignment. ***Across both paradigms, the generator's own capacity is decoupled from the evaluation phase, structurally omitting a fundamental source of alignment signal.***

We close this gap by introducing **GenRM-Flow**, a reward modeling framework in which the video generator itself serves as the process-aware reward model. The starting observation is a structural property of offline preference learning: any preference optimization designed to approximate a preference dataset mathematically targets a strictly defined KL-constrained optimal policy. By isolating the reward term from this analytical solution and utilizing the flow-matching velocity-prediction error as a surrogate for generation log-likelihood, a unified empirical reward formulation emerges. Because these preference objectives structurally target the identical optimal density shift, the implicit reward is consistently extracted directly from the generator's continuous-time denoising errors. After training, this evaluation is computed from the same velocity predictions the generator already produces, contrasted against a frozen reference model. Three structural properties fall out of this construction rather than requiring engineering. First, it operates natively in the noisy latent space without VAE decoding bottlenecks. Second, it is intrinsically process-aware, as the continuous-time integral is mathematically decomposable across any integration step. Most distinctively, the reward signal is sourced purely from the model's *generative capacity*, requiring no auxiliary prediction head or separate evaluation server.

We validate GenRM-Flow in two ways. First, the resulting reward model achieves strong discriminative accuracy on held-out preference pairs, outperforming both VLM-based and feature-extractor-based reward models, with a consistent margin over the latter across all noise levels. Second, the reward model is operationally effective for online policy improvement. Plugged into FlowAWR [4] as the per-step process reward source, we demonstrate that by substituting true data with generated end-states as trajectory proxy targets and applying group-wise Softmax normalization, under which the intractable partition function cancels, the offline-derived reward signal successfully drives measurable downstream generation improvement. **We summarize our contributions as follows:**

- We propose **GenRM-Flow**, a reward modeling framework that closes a structural gap in current video reward modeling: where prior approaches score generations without using

generation, GenRM-Flow trains a video generator to serve as a process-aware reward model through preference optimization alone, with no auxiliary head.

- We validate the operational efficacy of the resulting reward model by integrating it as a per-step process reward in FlowAWR. We show that with group-wise reward normalization and trajectory proxy targets, this generation-based reward successfully drives sustained online policy improvement.

2 Related Work

Reward Models for Video Generation. Reward modeling is now central to aligning generative models [19, 21, 22, 37, 42, 43, 62, 66–69] with human preferences. Existing video reward models are predominantly outcome-based [47], scoring fully generated outputs: T2VQA [17] and VideoScore [8] regress human ratings, VideoAlign [28] trains on pairwise preferences via the Bradley-Terry loss [2, 45], and others adapt VLMs [38, 65, 72] or add chain-of-thought reasoning for interpretability [9, 58, 59, 61, 63]. Operating only on near-complete outputs, these models provide no guidance during the early denoising stages that determine fundamental video attributes.

Process reward models [24, 55, 75] address this gap in the language domain by supervising intermediate steps. For visual generation, LPO [73] pioneered using the diffusion model itself as a noise-aware latent reward for step-level preference optimization in images, and PAVRM [36] extends this to video by attaching a separate reward head atop frozen VGM features. PAVRM, however, treats the generator as a feature extractor: the reward comes from the auxiliary head rather than the model's own velocity predictions, leaving the generative capacity unused in scoring. In contrast, we derive a fully implicit process reward from the policy model's prediction error relative to the reference, requiring no auxiliary parameters and aligning with the generation process at every denoising timestep.

Reinforcement Learning for Visual Generation. The integration of reinforcement learning (RL) into large language models [6, 11–13, 40] and multimodal LLMs has significantly advanced reasoning capabilities [49, 56, 57, 74], driven by the shift from replicating training data to refining reasoning trajectories through reward optimization. Proximal Policy Optimization (PPO) [32, 46] initiated this paradigm, while Group Relative Policy Optimization (GRPO) [3, 48] simplified advantage computation by eliminating the critic network. GRPO has been successfully applied to visual understanding [20, 26, 30, 33, 50, 70, 71, 76] and, more recently, to reward modeling for visual generation [9, 23, 64]. We provide a per-step process reward derived from the VGM's own predictions, which addresses the sparse reward problem in RL-based video generation fine-tuning by supplying dense supervision at every denoising timestep.

3 Method

3.1 Preliminaries

Rectified flow and flow matching. Recent video generation models [5, 16, 54] operate in rectified flow [25, 29], which establishes a continuous transport between data and noise distributions. Given clean data $\mathbf{x}_0 \sim p_{\text{data}}(\cdot \mid c)$ and noise $\mathbf{x}_1 \sim \mathcal{N}(0, I)$, the framework defines the interpolation

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

and trains a velocity field $\mathbf{v}_\theta(\mathbf{x}_t, c, t)$ with

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, c, t) - \mathbf{v}\|^2\right], \qquad \mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0. \tag{2}$$

We write $\mathcal{L}_\theta(\mathbf{x}_t, c) := \|\mathbf{v}_\theta(\mathbf{x}_t, c, t) - \mathbf{v}\|^2$ for the per-sample velocity error throughout. We will rely on a single property of this loss: it is a Monte Carlo surrogate for the negative log-likelihood of $\mathbf{x}_0$ under model θ [14, 15]. Up to a positive timestep weighting,

$$-\log p_\theta(\mathbf{x}_0 \mid c) \;\lesssim\; \mathbb{E}_{t,\mathbf{x}_1}[\mathcal{L}_\theta(\mathbf{x}_t, c)] + \text{const}. \tag{3}$$

Eq. (3) lets us treat $-\mathcal{L}_\theta$ as a stand-in for $\log \pi_\theta$, transporting any quantity defined in terms of policy log-likelihood directly into the flow matching setting.

3.2 Generation as Reward via Preference Optimization

The starting point is a structural observation about offline preference optimization. Given a static preference dataset $\mathcal{D}^D = \{(c, \mathbf{x}_w, \mathbf{x}_l)\}$ ranked along a target dimension D, a broad spectrum of fine-tuning objectives fundamentally optimize the generator to approximate a strictly defined optimal offline policy. These objectives include both contrastive methods, such as DPO [44] and IPO [1], and non-contrastive approaches like reward-weighted regression (RWR) [28]. Regardless of their specific loss formulations, these methods implicitly or explicitly target the analytical solution of a KL-constrained reward maximization problem (the formal derivation is detailed in Appendix A.1):

$$\pi^*(\mathbf{x}|c) = \frac{1}{Z(c)}\pi_{\text{ref}}(\mathbf{x}|c)\exp\left(\frac{R^D(\mathbf{x}, c)}{\beta}\right), \tag{4}$$

where $Z(c)$ is the partition function and β controls the KL-divergence penalty. Because the optimal policy π^* is mathematically defined by scaling the reference distribution π_{ref} via the exponentiated reward, updating the policy parameters θ to approximate π^* inherently forces the generator to internalize the preference signal. Consequently, an implicit scalar function $\hat{r}_{\mathcal{F}}(\mathbf{x}, c; \theta, \theta_{\text{ref}})$ emerges directly from the density shift dictated by the chosen objective $\mathcal{F}$. Minimizing the respective loss is mathematically equivalent to calibrating this implicit function on the static dataset: the optimization provably fits θ such that $\hat{r}_{\mathcal{F}}$ evaluates $\mathbf{x}_w$ systematically higher than $\mathbf{x}_l$.

Consequently, preference optimization achieves a dual objective without architectural modification. Explicitly, it aligns the generative distribution to favor D-preferred outputs during sampling. This alignment process structurally configures the scalar function $\hat{r}_{\mathcal{F}}$ to numerically rank inputs along dimension D, which aligns with the function of a reward model. Upon the completion of training, this scalar function is implicitly contained within the updated weights, accessible via a standard forward pass, which forms our central claim: ***A video generation model trained on preference data inherently functions as a reward model.*** Figure 2 contrasts this generation-based paradigm with existing representation-based reward models.

Three structural properties substantiate this claim. First, the reward derives directly from *generative capacity* rather than intermediate feature extraction, for $\hat{r}_{\mathcal{F}}$ is a strictly defined function of the generation likelihood (or the velocity-prediction error $\mathcal{L}_\theta$ via Eq. (3)). Second, the framework requires no architectural modification or auxiliary prediction heads. The trained policy $\theta^\star$, whether fully fine-tuned or adapted via LoRA, maintains exact topological equivalence with the base generator, computing scalar evaluations algebraically from native tensor outputs. Third, it eliminates the infrastructure overhead of independent reward modeling. By repurposing the generative computational graph, generation and evaluation are unified within the same system instance, requiring no separate reward server or standalone discriminator network.

3.3 Empirical Reward via Velocity-Prediction Discrepancy

We formalize this structural equivalence algebraically by instantiating the theoretical reward within the flow matching framework. By isolating the reward term from Eq. (4), the theoretical reward function is mathematically guaranteed to be:

$$R^D(\mathbf{x}, c) = \beta \log \frac{\pi^*(\mathbf{x}|c)}{\pi_{\text{ref}}(\mathbf{x}|c)} + \beta \log Z(c). \tag{5}$$

The partition function $Z(c)$ introduces an intractable, prompt-specific baseline offset. Because $Z(c)$ is strictly dependent on the condition c and invariant to $\mathbf{x}$, it is neutralized in downstream applications (e.g., via relative ranking or explicit baseline subtraction, as detailed in Section 3.4). Consequently, the operative component of the reward is isolated strictly to the log-likelihood ratio.

Assuming the preference optimization successfully drives the parameterized policy $\pi_{\theta^\star}$ to approximate π^*, we approximate the exact generation log-likelihood using its tractable flow-matching surrogate via the inequality established in Eq. (3). By treating $-\mathcal{L}_\theta$ as a bounded proxy for $\log \pi_\theta$, an empirical reward emerges as the expected difference in velocity-prediction errors. Note that this formulation is an approximation subject to the tightness of the variational bound:

$$\hat{r}(\mathbf{x}, c; \theta^\star, \theta_{\text{ref}}) \;\approx\; \beta\, \mathbb{E}_{t,\, \mathbf{x}_1}[\mathcal{L}_{\text{ref}}(\mathbf{x}_t, c) - \mathcal{L}_{\theta^\star}(\mathbf{x}_t, c)] + \beta \log Z(c). \tag{6}$$

It exposes the operational mechanism: a sample receives a higher scalar evaluation if the trained model $\theta^\star$ exhibits a lower denoising error than the reference model θ_{ref}, conditioned on the assumption that the error divergence proportionally reflects the preference density shift.

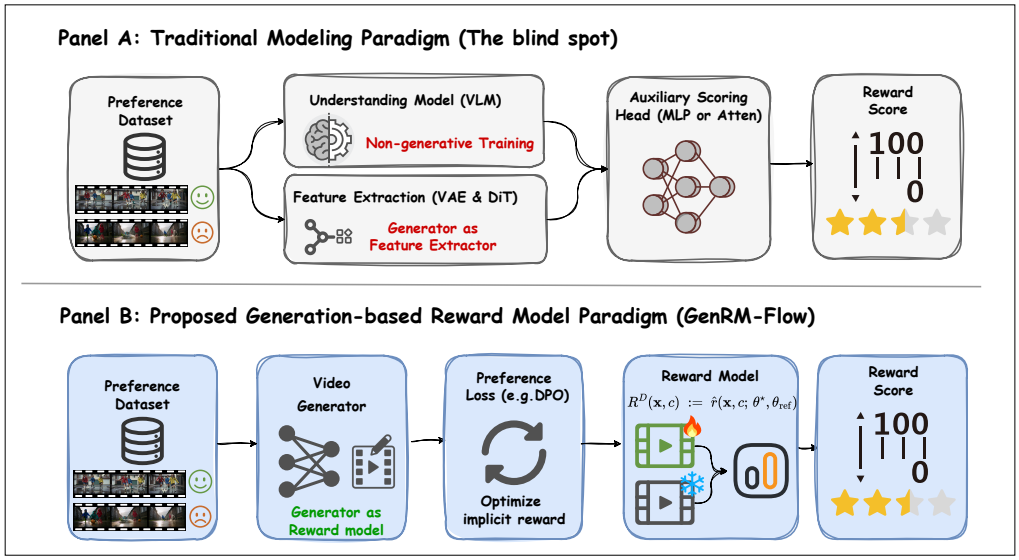


Figure 2: The pipeline of **GenRM-Flow**. **(Panel A)** Traditional paradigms rely on *representational capacity*, utilizing a frozen backbone (e.g., a VLM or base generator) as a feature extractor coupled with an auxiliary discriminative head. **(Panel B)** The proposed framework sources the reward directly from *generative capacity*. Following preference optimization, the implicit reward $R^D(\mathbf{x}, c)$ is algebraically extracted from the velocity-prediction errors. The scalar score is computed natively by contrasting the active policy (flame, $\theta^\star$) against a frozen reference (snowflake, θ_{ref}), requiring no architectural modifications or auxiliary heads.

Generality Across Preference Objectives. This theoretical derivation establishes that the reward evaluation is structurally defined by the underlying fine-tuning objective $\mathcal{F}$. Specific preference optimization algorithms yield specific evaluations corresponding to their density-shift mechanisms (summarized in Table 1). Contrastive objectives, such as DPO [44] and IPO [1], inherently train $\theta^\star$ against a base distribution, yielding the relative evaluation paradigm defined in Eq. (6). Alternatively, non-contrastive objectives like RWR [28] operate via preferred-only likelihood maximization, bypassing the reference policy entirely to yield an absolute (reference-free) reward evaluation of the form $R^D(\mathbf{x}, c) \propto -\mathbb{E}_{t,\mathbf{x}_1}[\mathcal{L}_{\theta^\star}(\mathbf{x}_t, c)]$; in practice, we instantiate RWR with binary weights (one for preferred samples and zero otherwise), which reduces the weighted objective to likelihood maximization on preferred samples. We provide rigorous mathematical proofs establishing how both DPO and RWR theoretically originate from the optimal policy π^* within the rectified flow framework in Appendix A.2. Regardless of the specific $\mathcal{F}$, the evaluation strictly relies on the respective generative objective, yielding a unified operational pathway:

$$\underbrace{\mathcal{F} \text{ on } \mathcal{D}^D}_{\text{preference optimization}} \xrightarrow{\text{minimize}} \underbrace{\theta^\star = \arg\min_\theta \mathcal{L}_{\mathcal{F}}(\theta)}_{\text{aligned generator}} \xRightarrow{\text{induces}} \underbrace{R^D(\mathbf{x}, c) := \hat{r}(\mathbf{x}, c; \theta^\star, \theta_{\text{ref}})}_{\text{implicit reward model}}. \quad (7)$$

Paradigm Differentiation. The mathematical formulation in Eq. (6) structurally distinguishes the resulting evaluation framework from existing VLM-based and feature-extractor-based paradigms through two operational properties. First, the reward evaluation is executed natively within the noisy latent space. Because the velocity-prediction error $\mathcal{L}_{\theta^\star}$ operates directly on the intermediate latent state $\mathbf{x}_t$, it systematically bypasses RGB pixel-space transformations. This eliminates the severe computational bottleneck of VAE decoding that constrains standard VLM scorers. Second, the evaluation is intrinsically process-aware. Because the expectation in Eq. (6) is formulated over continuous timesteps $t \in [0, 1]$, the integral is mathematically decomposable. Consequently, a single set of weights can output a calibrated evaluation score at any arbitrary point along the denoising trajectory, including the early high-noise regimes where macroscopic structure and motion dynamics form. While recent feature-extractor methods [36] must painstakingly engineer auxiliary prediction heads to approximate these capabilities, the proposed framework inherits them as a direct analytical consequence of the generative objective.

Table 1: Three preference fine-tuning objectives as optimization pathways to obtain the trained weights $\theta^\star$. While their loss formulations $\mathcal{L}_\mathcal{F}(\theta)$ vary significantly—ranging from bounded contrastive margins to reference-free likelihood maximization—they all structurally aim to approximate the optimal offline policy π^*. Consequently, during the evaluation phase, they implicitly define reward readouts corresponding to their theoretical likelihood shifts: the relative readout of Eq. (6) for contrastive objectives and a reference-free readout for RWR.

Fine-tuning Objective $\mathcal{F}$	Family	Flow Matching Loss Formulation $\mathcal{L}_\mathcal{F}(\theta)$
RWR [28]	Preferred-only likelihood	$\mathbb{E}_{t,\mathbf{x}_1}[\mathcal{L}_\theta(\mathbf{x}_{w,t})]$
IPO [1]	Identity-margin contrastive	$\mathbb{E}_{t,\mathbf{x}_1}\left[\left([(\mathcal{L}_{\text{ref}}-\mathcal{L}_\theta)\big\|_{\mathbf{x}_{w,t}}-(\mathcal{L}_{\text{ref}}-\mathcal{L}_\theta)\big\|_{\mathbf{x}_{l,t}}]-\frac{1}{2\tau}\right)^2\right]$
DPO [44]	BT contrastive (saturating)	$-\mathbb{E}_{t,\mathbf{x}_1}\left[\log\sigma\left(\beta\left[(\mathcal{L}_{\text{ref}}-\mathcal{L}_\theta)\big\|_{\mathbf{x}_{w,t}}-(\mathcal{L}_{\text{ref}}-\mathcal{L}_\theta)\big\|_{\mathbf{x}_{l,t}}\right]\right)\right]$

3.4 Process Reward Feedback Learning via FlowAWR

To verify the operational validity of the derived latent-space reward, we integrate it into FlowAWR [4]. While standard RLHF paradigms compute a global advantage using external VLM scorers after full VAE decoding, our framework bypasses this bottleneck by directly extracting a process-aware latent advantage $A^t(\mathbf{x}_t|c)$.

During rollout, the sampling policy θ_{old} generates G trajectories conditioned on prompt c with a deterministic ODE solver. To compute the intermediate reward, we adopt the policy's final generated latent $\hat{\mathbf{x}}_0^{(i)}$ as the proxy ground-truth target for its respective trajectory, and obtain $\mathbf{x}_t^{(i)}$ by forward-noising $\hat{\mathbf{x}}_0^{(i)}$ via Eq. (1) at the sampled timesteps. Crucially, the evaluation is performed by the offline-trained reward model $\theta^\star$ (e.g., the frozen DPO LoRA) and the reference θ_{ref}, which are decoupled from the active generation process. The raw process reward at timestep t is evaluated:

$$r_t(\mathbf{x}_t^{(i)}, c) \;=\; \beta\Big[\mathcal{L}_{\text{ref}}(\mathbf{x}_t^{(i)}, \hat{\mathbf{x}}_0^{(i)}, c) - \mathcal{L}_{\theta^\star}(\mathbf{x}_t^{(i)}, \hat{\mathbf{x}}_0^{(i)}, c)\Big]\,. \tag{8}$$

Because $\theta^\star$ acts strictly as an independent evaluator during the RL phase, substituting the true data with the generated end-state $\hat{\mathbf{x}}_0^{(i)}$ provides a valid velocity target to measure the trajectory's alignment quality, without introducing self-evaluation bias. Furthermore, Eq. (8) represents only the log-likelihood ratio and omits the intractable partition function $\beta \log Z(c)$. Since this term is shared by all trajectories sampled for the same prompt c, it cancels under any group-relative normalization. We adopt a Softmax transformation over the group:

$$\tilde{r}_t(\mathbf{x}_t^{(i)}, c) = \frac{\exp\Big(r_t(\mathbf{x}_t^{(i)}, c)\Big)}{\sum_{j=1}^{G}\exp\Big(r_t(\mathbf{x}_t^{(j)}, c)\Big)}. \tag{9}$$

Beyond removing the prompt-dependent offset, this transformation bounds the reward within $[0, 1]$ and, through its exponential form, emphasizes the top-ranked trajectories in each group, converting the reward signal into relative rankings in the spirit of recent preference-based paradigms such as Pref-GRPO [60].

The timestep-specific advantage $A^t(\mathbf{x}_t^{(i)}|c) \;=\; G\,\tilde{r}_t(\mathbf{x}_t^{(i)}, c) - 1$ is subsequently computed from these Softmax scores, following the group-based advantage estimate of FlowAWR. The policy is then updated by regressing $\mathbf{v}_\theta(\mathbf{x}_t^{(i)}, c, t)$ toward the advantage-weighted rectified target $\mathbf{v}_{\theta_{\text{old}}} + A^t(\mathbf{x}_t^{(i)}|c)\,(\mathbf{v}^{(i)} - \mathbf{v}_{\theta_{\text{old}}})$, where $\mathbf{v}_{\theta_{\text{old}}} = \mathbf{v}_{\theta_{\text{old}}}(\mathbf{x}_t^{(i)}, c, t)$ and $\mathbf{v}^{(i)} = \mathbf{x}_1^{(i)} - \hat{\mathbf{x}}_0^{(i)}$. Trajectory sampling and the remaining optimization settings strictly follow FlowAWR defaults, isolating the efficacy of the offline-derived process reward in driving online optimization.

4 Experiments

4.1 Experimental Setup

Dataset. We collect **7,170** text prompts from the Internet, spanning multiple semantic categories including *people, animals, scenery, sports, objects, architecture, food*, and *vehicles*. To construct

Table 2: Statistics of the collected video pool. We aggregate 48,757 videos generated by ten text-to-video models spanning both proprietary (Kling-3.0/1.6, Seedance-1.0, Sora, Pika-2.2) and open-source (StepVideo-T2V, Wan2.2/2.1, Ruyi, LTX-Video-0.9.5) systems, balanced across sources to prevent any single model from dominating the preference distribution. Per-source statistics include FPS, resolution, clip duration, sample count, and proportion.

T2V Model	Open Source	FPS	Resolution	Duration	Num	Proportion
Kling-3.0 [19]	N	24.0	1280*720	5.0s	5127	10.52%
Kling-1.6 [18]	N	24.0	1280*720	5.0s	4893	10.04%
Seedance-1.0 [5]	N	24.0	1248*704	5.0s	5212	10.69%
Sora [39]	N	30.0	1280*720	4.0s	4586	9.41%
Pika-2.2 [41]	N	24.0	1280*720	5.0s	4692	9.62%
StepVideo-T2V [34]	Y	25.0	992*544	5.0s	5043	10.34%
Wan2.2 (14B) [54]	Y	16.0	832*480	5.0s	4951	10.15%
Ruyi [51]	Y	24.0	1008*576	5.0s	4728	9.70%
Wan2.1 (1.3B) [54]	Y	16.0	832*480	5.0s	5032	10.32%
LTX-Video-0.9.5 [7]	Y	25.0	704*480	4.8s	4493	9.21%
All: 48757 videos						

preference data, we employ 10 text-to-video generation models (Table 2); for each prompt, 6 to 8 models are randomly selected to generate videos, producing **48,757** videos in total. All generated videos are resized to a unified resolution of 832×480 at 16 fps with 5 s duration. 40 trained annotators then view pairs of videos generated from the same prompt and record pairwise preferences based on overall motion quality. The annotation protocol evaluates three dimensions (Table 3): (i) anatomical correctness, (ii) spatial plausibility, and (iii) temporal coherence. After filtering low-agreement annotations and rebalancing across prompt categories and model pairs, we obtain **84,597** preference pairs. We randomly hold out **4,800** pairs for evaluation; the remaining **79,797** pairs form the training set $\mathcal{D}$. All methods are trained and evaluated on the same splits to ensure a fair comparison. More details on the dataset are provided in Appendix C.

Training Details and Baselines. We employ the Wan2.1-T2V family [54] as our foundation models, validating our method on both the 1.3B and 14B variants. For our method, we use the AdamW optimizer [31] with a learning rate of 1×10^{-5} and LoRA rank 64. The DPO coefficient β (Table 1) is set to 2000. During FlowAWR training, we set $\beta = 1$ in the process reward of Eq. (8); since this reward enters the Softmax in Eq. (9), β effectively acts as its inverse temperature. Meanwhile, we benchmark our method against two baselines: VideoAlign [28] and PAVRM [36]. For VideoAlign [28], we use Qwen3-VL-8B as the base model. All videos are sampled at 16 fps with a full resolution of 832×480 pixels. For PAVRM [36], the original paper trains only on the 14B backbone; we extend it to both 1.3B and 14B using the same architecture: the first 8 DiT blocks serve as the spatiotemporal feature extractor, followed by a query-based attention aggregation layer and a three-layer MLP. The feature dimension is 5120 for the 14B model and 1536 for 1.3B.

Evaluation. We evaluate from two perspectives. For reward model evaluation, we measure pairwise preference prediction accuracy on the held-out 4,800 pairs, reported per timestep bin. To further validate the effectiveness of the reward model, we adopt each reward model as the reward function to train Wan2.1-T2V using FlowAWR [4], and assess the trained models with two protocols: (1) **Generation metrics**: following VBench [10], we report motion smoothness (MS), dynamic degree (DD), subject consistency (SC), and human action (HA) on both our inner test set and the out-of-domain VBench prompt suite. (2) **HPSv3** [35]: we track HPSv3 as an external visual-quality metric throughout FlowAWR training. Following [52], we score all sampled frames and report the average of the top 30% scoring frames, which alleviates the effect of low scores induced by content inconsistency across temporal transitions.

4.2 Main Results

Pairwise Preference Accuracy. We compare the pairwise preference prediction accuracy of our implicit reward formulation against PAVRM [36] and VideoAlign [28]. As shown in Table 4, on the 14B backbone, our method reaches 88.96% average accuracy, exceeding PAVRM (82.64%)

Table 3: Evaluation dimensions and assessment criteria used in the human annotation framework for pairwise preference labeling.

Evaluation Dimension	Definition and Assessment Criteria
Anatomical Correctness	Evaluates whether humans or animals exhibit **abnormal anatomical deformations or physiological anomalies**. - **Proportional Integrity:** Assesses whether body parts including limbs, face, head, neck, and torso maintain correct anatomical proportions and configurations. - **Deformation Artifacts:** Identifies abnormal twisting, bending, misalignment, quantity anomalies, or fusion of body parts.
Spatial Plausibility	Evaluates whether **spatial relationships and physical interactions** between objects within each frame comply with physical laws. - **Contact Plausibility:** Assesses whether collisions and contact between objects follow realistic physical logic. - **Structural Integrity:** Identifies distortions, collapses, or deformations that violate physical constraints. - **Penetration Artifacts:** Detects instances of objects passing through each other without physical interaction. - **Trajectory Consistency:** Evaluates whether object motion trajectories adhere to physical laws.
Temporal Coherence	Evaluates the **stability and continuity** of subjects across consecutive frames. - **Attribute Stability:** Color and texture of subjects should not fluctuate without cause. - **Existence Stability:** Subjects should not appear or disappear spontaneously; in multi-person scenes, the number of individuals should remain constant. - **Morphological Stability:** Limbs should not flash, fragment, disappear, or change in quantity during motion. - **Action Continuity:** Motion transitions should be coherent without repetitive flashing or abrupt jumps.

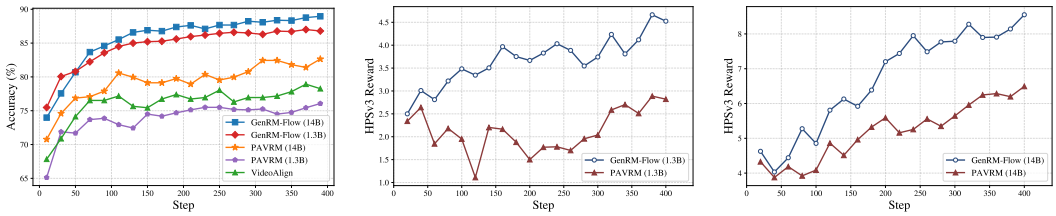


Figure 3: (Left) Pairwise preference prediction accuracy during training. (Middle, Right) HPSv3 reward trajectories during FlowAWR training on the 1.3B and 14B backbones. The implicit process reward yields higher sustained HPSv3 evaluation scores compared to PAVRM.

by 6.32 points; on the 1.3B backbone, it reaches 86.79% versus 76.06%, a 10.73-point gap. Both configurations also surpass VideoAlign (78.26%). The advantage persists across all five uniformly spaced timestep bins over $t \in [0, 1]$, and remains pronounced in the high-noise regime, where our method sustains at least 81.00% while PAVRM drops to 78.02% (14B) and 71.25% (1.3B). Figure 3 (left) tracks preference accuracy during training: our method converges faster and to a higher terminal value than all baselines on both scales. This suggests that the relative prediction error between policy and reference models provides a stronger preference signal than PAVRM's independent feature-extractor design, while introducing no auxiliary reward-model parameters.

Generation Quality with FlowAWR. To assess whether reward-model accuracy translates into generation gains, we drive FlowAWR fine-tuning with our process-level rewards (PAVRM and GenRM-Flow) under shared hyperparameters; comparisons against the sequence-level VideoAlign [28]

Table 4: Pairwise preference prediction accuracy (%) on the held-out evaluation set. The performance of our implicit process reward is compared against baseline reward models (VideoAlign and PAVRM) across five uniformly spaced noise timestep intervals $t \in [0, 1]$. **Bold** denotes the highest accuracy per column within each backbone scale.

Base Model	Method	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]	Avg
Qwen3-VL-8B	VideoAlign [28]	–	–	–	–	–	78.26
Wan2.1-T2V-1.3B	PAVRM [36]	77.19	77.50	78.27	76.11	71.25	76.06
	GenRM-Flow	**87.77**	**88.08**	**88.49**	**88.61**	**81.00**	**86.79**
Wan2.1-T2V-14B	PAVRM [36]	81.87	83.85	84.79	84.69	78.02	82.64
	GenRM-Flow	**89.10**	**89.94**	**90.46**	**90.75**	**84.55**	**88.96**



Figure 4: Qualitative comparison between PAVRM and **GenRM-Flow** as the per-step reward signal in FlowAWR with a 1.3B backbone. Both runs use FlowAWR with the same configuration, differing only in the reward signal. Red boxes mark localized structural deformations, spatial inconsistencies, and temporal artifacts produced by PAVRM. Across scenes, our reward yields visibly fewer geometric defects and better spatial and temporal continuity. More visualization results are shown in Figure 5 (Appendix B).

baseline are deferred to Appendix D. **Quantitative results.** Table 5 reports VBench scores. Our reward yields the best results across all four metrics on both backbone scales, whereas PAVRM yields only marginal gains on the 1.3B model and even slightly lowers VBench Subject Consistency ($97.56 \rightarrow 97.52$), mirroring the prediction-accuracy gap in Table 4. **Qualitative analysis.** Figure 3 (middle, right) tracks HPSv3 during fine-tuning: ours stays consistently higher, while PAVRM training is lower and more volatile. Qualitatively (Figure 4), PAVRM outputs exhibit localized defects, including limb distortion during paragliding and implausible object interactions, while ours produce anatomically and spatially coherent videos. Prompts are listed in Appendix E.

4.3 Ablation Studies

Intra-group Reward Normalization. We study how the raw process reward r_t (Eq. (8)) should be normalized within each group during FlowAWR training, comparing three schemes on Wan2.1-T2V-1.3B: the raw reward without group normalization, Z-score normalization, and Softmax (Eq. (9)). As shown in Table 6, training with the raw reward collapses within 250 steps, which we attribute to the large scale variation of r_t across prompts and timesteps in the absence of group normalization. Both Z-score and Softmax remove the prompt-dependent offset and stabilize training. Softmax further

Table 5: Comprehensive quantitative comparison of process-guided generation on our inner test set and the out-of-domain VBench benchmark. **Bold** denotes the highest metric value per backbone.

Base Model	Method	Inner Test Set				VBench [10]			
		MS ↑	DD↑	SC↑	HA↑	MS↑	DD↑	SC↑	HA↑
Wan2.1-T2V-1.3B	Pretrain	98.31	59.07	97.02	88.15	98.52	65.19	97.56	94.00
	PAVRM [36]	98.56	63.59	97.18	89.73	98.71	67.83	97.52	94.68
	GenRM-Flow	**98.88**	**72.43**	**98.28**	**93.16**	**98.95**	**73.46**	**97.70**	**96.25**
Wan2.1-T2V-14B	Pretrain	98.12	61.18	97.18	89.32	98.30	65.46	97.52	95.40
	PAVRM [36]	98.34	68.59	97.82	91.45	98.41	71.82	97.68	96.18
	GenRM-Flow	**98.76**	**77.48**	**98.45**	**95.08**	**99.03**	**74.87**	**97.73**	**97.73**

outperforms Z-score on all metrics (e.g., raising VBench Dynamic Degree from 69.85 to 73.46), which we attribute to its bounded, nonlinear emphasis on the top-ranked trajectories in each group.

Table 6: Ablation on intra-group normalization mechanisms during FlowAWR training (Wan2.1-T2V-1.3B, evaluated on VBench). "Policy Collapse" indicates training collapse. **Bold** denotes the best results among all.

Normalization Mechanism	Training Status	MS↑	DD↑	SC↑	HA↑
Raw r_t (no normalization)	Policy Collapse (∼250 steps)	N/A	N/A	N/A	N/A
Z-score Normalization	Converged	98.74	69.85	97.61	95.61
Softmax (Ours)	Converged	**98.95**	**73.46**	**97.70**	**96.25**

5 Conclusion

We have introduced **GenRM-Flow**, a reward modeling framework in which a video generator, trained on preference data through standard preference optimization, serves as a process-aware reward model in flow matching. The reward signal is sourced from the generator's own generative capacity rather than from an external scorer or an auxiliary head, and is computed from the velocity predictions of the trained and reference models at any timestep along the denoising trajectory. As a per-step process reward source in FlowAWR, the resulting reward model drives measurable downstream generation improvement.

Limitations and future work. GenRM-Flow inherits the trajectory-mismatch issue common to RL fine-tuning of flow matching models: the discrete-time scheduler used during RL rollout differs from the inference solver, and the resulting distribution shift can dampen the quality of the per-step reward signal. Resolving this within a continuous-time treatment of process reward, rather than through trajectory shortcuts, is a natural next step.

Acknowledgments and Disclosure of Funding

This work received no third-party funding. The authors declare no competing interests.

References

[1] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[3] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

[4] Zheming Fu, Ruizhe He, Wei Shang, Xiaoxiao Ma, Lei Wang, Chang Liu, and Siming Fu. FlowAWR: Online adaptive flow reinforcement via advantage-weighted rectification. *arXiv preprint arXiv:2606.30376*, 2026.

[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.

[6] Google. Gemini-2.5-pro. https://deepmind.google/models/gemini/pro/, 2025.

[7] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[8] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan 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.

[9] 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. *arXiv preprint arXiv:2509.22799*, 2025.

[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.

[11] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[13] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[14] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36:65484–65516, 2023.

[15] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[16] 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.

[17] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7793–7802, 2024.

[18] Kuaishou. Kling-1.6. https://app.klingai.com/global, 2025.

[19] Kuaishou. Kling-3.0. https://app.klingai.com/global, 2025.

[20] Jinghan Li, Junfeng Fang, Jinda Lu, Yuan Wang, Xiaoyan Guo, Tianyu Zhang, Xiang Wang, and Xiangnan He. Enhancing multi-modal llms reasoning via difficulty-aware group normalization. *arXiv preprint arXiv:2602.21743*, 2026.

[21] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iqAFhWistW.

[22] Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. SPEED: Scalable, precise, and efficient concept erasure for diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[23] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025.

[24] 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. URL https://openreview.net/forum?id=v8L0pN6EOi.

[25] 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. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[26] Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. Diversegrpo: Mitigating mode collapse in image generation via diversity-aware grpo. *arXiv preprint arXiv:2512.21514*, 2025.

[27] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

[28] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[29] 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. URL https://openreview.net/forum?id=XVjTT1nw5z.

[30] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Jinda Lu, Jinghan Li, Yuan Gao, Junkang Wu, Jiancan Wu, Xiang Wang, and Xiangnan He. Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization, 2025.

[33] Jinda Lu, Junkang Wu, Jinghan Li, Xiaojun Jia, Shuo Wang, Yifan Zhang, Junfeng Fang, Xiang Wang, and Xiangnan He. Dama: Data-and model-aware alignment of multi-modal llms. In *International Conference on Machine Learning*, pages 40726–40740. PMLR, 2025.

[34] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Haonan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jiahao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. URL https://arxiv.org/abs/2502.10248.

[35] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

[36] Xiaoyue Mi, Wenqing Yu, Jiesong Lian, Shibo Jie, Ruizhe Zhong, Zijun Liu, Guozhen Zhang, Zixiang Zhou, Zhiyong Xu, Yuan Zhou, et al. Video generation models are good latent reward models. *arXiv preprint arXiv:2511.21541*, 2025.

[37] MiniMax. Hailuo. https://hailuoai.com/, 2024.

[38] 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.

[39] OpenAI. Openai sora. https://openai.com/sora, 2024.

[40] OpenAI. Gpt-5. https://openai.com/index/introducing-gpt-5/, 2025.

[41] Pika-Labs. Pika-v2.2. https://pikalabs.org/pika-2-2, 2025.

[42] Junxiang Qiu, Lin Liu, Shuo Wang, Jinda Lu, Kezhou Chen, and Yanbin Hao. Accelerating diffusion transformer via gradient-optimized cache. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17608–17617, 2025.

[43] Junxiang Qiu, Shuo Wang, Jinda Lu, Lin Liu, Houcheng Jiang, Xingyu Zhu, and Yanbin Hao. Accelerating diffusion transformer via error-optimized cache. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9588–9597, 2025.

[44] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[45] Pejaver V Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[47] 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.

[48] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

[49] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable r1-style large vision-language model. *CoRR*, abs/2504.07615, 2025.

[50] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, 2024.

[51] CreateAI Team. Ruyi-mini-7b. https://github.com/IamCreateAI/Ruyi-Models, 2024.

[52] Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

[53] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

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

[55] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: 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.

[56] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025.

[57] Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. Llava-critic-r1: Your critic model is secretly a strong policy model. *CoRR*, abs/2509.00676, 2025.

[58] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024.

[59] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multi-modal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025.

[60] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025.

[61] 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.

[62] Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28759–28768. IEEE, 2025.

[63] Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025.

[64] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025.

[65] 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.

[66] Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. Diffusion models for generative outfit recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 1350–1359, 2024.

[67] Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, pages 264–274, 2025.

[68] Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. Personalized generation in large model era: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24607–24649, 2025.

[69] Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. Drc: Enhancing personalized image generation via disentangled representation composition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9667–9676, 2025.

[70] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

[71] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv e-prints*, pages arXiv–2405, 2024.

[72] 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.

[73] 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.

[74] Yifan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, and Liang Wang. R1-reward: Training multimodal reward model through stable reinforcement learning. *CoRR*, abs/2505.02835, 2025.

[75] 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.

[76] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

A Extended Theoretical Derivations and Proofs

A.1 Analytical Derivation of the Optimal Policy

This section derives the closed-form optimal solution π^* for the preference optimization objective under a global KL-divergence constraint. In conditional generation tasks, the policy $\pi(\cdot|c)$ aims to maximize the expected implicit reward $R(\mathbf{x}, c)$ while constraining its divergence from a reference distribution $\pi_{\text{ref}}(\cdot|c)$. Given the marginal distribution $p(c)$ of the condition c, the global optimization problem is formulated as:

$$\max_{\pi} \mathbb{E}_{c\sim p(c)} \left[\mathbb{E}_{\mathbf{x}\sim\pi(\cdot|c)}[R(\mathbf{x}, c)] - \beta \mathbb{KL}(\pi(\cdot|c) \parallel \pi_{\text{ref}}(\cdot|c)) \right].$$

Because the marginal distribution $p(c)$ is an extrinsic constant defined by the dataset, maximizing the global expectation is mathematically equivalent to solving for the local functional extremum for each independent condition c. Expanding the local objective function into an integral form over the continuous variable space $\mathbf{x}$ yields:

$$J_c(\pi) = \int \pi(\mathbf{x}|c) R(\mathbf{x}, c) \mathrm{d}\mathbf{x} - \beta \int \pi(\mathbf{x}|c) \log \frac{\pi(\mathbf{x}|c)}{\pi_{\text{ref}}(\mathbf{x}|c)} \mathrm{d}\mathbf{x}.$$

To ensure that $\pi(\mathbf{x}|c)$ constitutes a valid probability density function for any c, the equality constraint $\int \pi(\mathbf{x}|c)\mathrm{d}\mathbf{x} = 1$ must be strictly enforced. By introducing a condition-dependent Lagrange multiplier $\mu(c)$, the corresponding Lagrangian functional is constructed:

$$\mathcal{L}(\pi, \mu) = J_c(\pi) + \mu(c) \left(1 - \int \pi(\mathbf{x}|c) \mathrm{d}\mathbf{x} \right).$$

To determine the extremum of this functional, we compute its functional derivative (the first variation) with respect to the probability density $\pi(\mathbf{x}|c)$ and set it to zero:

$$\frac{\delta \mathcal{L}}{\delta \pi(\mathbf{x}|c)} = R(\mathbf{x}, c) - \beta \left(\log \frac{\pi(\mathbf{x}|c)}{\pi_{\text{ref}}(\mathbf{x}|c)} + 1 \right) - \mu(c) = 0.$$

Algebraic rearrangement to isolate the logarithmic density term yields:

$$\log \pi(\mathbf{x}|c) = \log \pi_{\text{ref}}(\mathbf{x}|c) + \frac{R(\mathbf{x}, c)}{\beta} - \left(1 + \frac{\mu(c)}{\beta} \right).$$

Exponentiating both sides provides the explicit expression for the probability density:

$$\pi(\mathbf{x}|c) = \pi_{\text{ref}}(\mathbf{x}|c) \exp\left(\frac{R(\mathbf{x}, c)}{\beta} \right) \exp\left(-1 - \frac{\mu(c)}{\beta} \right).$$

The term $\exp(-1 - \mu(c)/\beta)$ is strictly independent of the generative variable $\mathbf{x}$. To satisfy the integral normalization condition of the probability density, the reciprocal of this constant term must act as the partition function $Z(c)$:

$$Z(c) = \exp\left(1 + \frac{\mu(c)}{\beta} \right) = \int \pi_{\text{ref}}(\mathbf{x}|c) \exp\left(\frac{R(\mathbf{x}, c)}{\beta} \right) \mathrm{d}\mathbf{x}.$$

Substituting $Z(c)$ to eliminate the Lagrange multiplier term yields the analytical solution for the optimal policy, verifying the formalization established in the main text (Eq. 4):

$$\pi^*(\mathbf{x}|c) = \frac{1}{Z(c)} \pi_{\text{ref}}(\mathbf{x}|c) \exp\left(\frac{R(\mathbf{x}, c)}{\beta} \right). \tag{10}$$

A.2 Theoretical Unification of RWR and DPO

As highlighted in the main text, both Reward-Weighted Regression (RWR) and Direct Preference Optimization (DPO) represent distinct alignment paradigms. Despite their divergent practical implementations—regression versus contrastive learning—both methods mathematically originate from the identical KL-constrained reward maximization framework. This section explicitly connects both objectives to the closed-form optimal density π^* derived in Appendix A.1.

A.2.1 Connection to RWR in Flow Matching

Under the rectified flow framework defined in the main text, the objective is to transport noise $\mathbf{x}_1 \sim \mathcal{N}(0, I)$ to a target data distribution. To align the model with the theoretical optimal policy π^*, the ideal Conditional Flow Matching (CFM) objective must sample the clean terminal states directly from π^*:

$$\mathcal{L}^*_{\text{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_1\sim\mathcal{N},\mathbf{x}_0\sim\pi^*(\cdot|c)}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, c, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right].$$

Because direct sampling from the optimal policy π^* is practically intractable prior to alignment, we apply importance sampling using the accessible reference policy π_{ref} (i.e., the pre-trained base model). This re-weights the expectation by the density ratio:

$$\mathcal{L}^*_{\text{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_1\sim\mathcal{N},\mathbf{x}_0\sim\pi_{\text{ref}}(\cdot|c)}\left[\frac{\pi^*(\mathbf{x}_0|c)}{\pi_{\text{ref}}(\mathbf{x}_0|c)}\|\mathbf{v}_\theta(\mathbf{x}_t, c, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right].$$

Substituting the density ratio with the analytical solution of the optimal policy yields:

$$\frac{\pi^*(\mathbf{x}_0|c)}{\pi_{\text{ref}}(\mathbf{x}_0|c)} = \frac{1}{Z(c)}\exp\left(\frac{R(\mathbf{x}_0, c)}{\beta}\right).$$

Because the partition function $Z(c)$ is independent of the optimized parameters θ, it serves strictly as a constant scaling factor and can be absorbed or discarded during gradient descent. Dropping $Z(c)$ rigorously recovers the classical RWR objective mapped into the continuous generation process:

$$\mathcal{L}_{\text{RWR}}(\theta) \propto \mathbb{E}_{t,\mathbf{x}_1,\mathbf{x}_0\sim\pi_{\text{ref}}}\left[\exp\left(\frac{R(\mathbf{x}_0, c)}{\beta}\right)\|\mathbf{v}_\theta(\mathbf{x}_t, c, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right]. \tag{11}$$

This derivation proves that exponentiated reward re-weighting emerges analytically from projecting the flow-matching velocity onto the unobservable optimal policy trajectory.

A.2.2 Connection to DPO

Unlike RWR's regression-based approximation, DPO bypasses explicit scalar reward modeling entirely. By algebraically rearranging the logarithmic form of the optimal policy, the unobserved theoretical reward $R(\mathbf{x}, c)$ can be reparameterized strictly in terms of policy probabilities:

$$R(\mathbf{x}, c) = \beta\log\frac{\pi^*(\mathbf{x}|c)}{\pi_{\text{ref}}(\mathbf{x}|c)} + \beta\log Z(c).$$

In the context of pairwise preference datasets $\mathcal{D} = \{(c, \mathbf{x}_w, \mathbf{x}_l)\}$, standard DPO models human preference using the Bradley-Terry (BT) model. The probability that generation $\mathbf{x}_w$ is preferred over $\mathbf{x}_l$ is dictated by the logistic function of their latent reward difference:

$$P(\mathbf{x}_w \succ \mathbf{x}_l \mid c) = \sigma(R(\mathbf{x}_w, c) - R(\mathbf{x}_l, c)).$$

Substituting the reparameterized reward into the reward difference elegantly cancels out the intractable partition function $Z(c)$, as it acts as a uniform baseline for a given condition c:

$$R(\mathbf{x}_w, c) - R(\mathbf{x}_l, c) = \beta\log\frac{\pi^*(\mathbf{x}_w|c)}{\pi_{\text{ref}}(\mathbf{x}_w|c)} - \beta\log\frac{\pi^*(\mathbf{x}_l|c)}{\pi_{\text{ref}}(\mathbf{x}_l|c)}.$$

By parameterizing the optimal policy π^* with a learnable neural network π_θ and minimizing the negative log-likelihood of the observed preference data, the exact DPO objective is isolated:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(c,\mathbf{x}_w,\mathbf{x}_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(\mathbf{x}_w|c)}{\pi_{\text{ref}}(\mathbf{x}_w|c)} - \beta\log\frac{\pi_\theta(\mathbf{x}_l|c)}{\pi_{\text{ref}}(\mathbf{x}_l|c)}\right)\right]. \tag{12}$$

This confirms that DPO structurally inherits the exact KL-constrained optimality criteria as RWR, differing only by reformulating the objective to optimize relative margins over pairwise comparisons rather than absolute weighted likelihoods. (Note: As detailed in Eq. 3 of the main text, mapping this exact likelihood objective into the flow-matching framework requires approximating $\log\pi_\theta$ with the velocity prediction error $-\mathcal{L}_\theta$).

B More Visualization Results

In this section, we provide additional qualitative examples to further demonstrate the efficacy of our proposed GenRM-Flow framework. Figure 5 presents an extended visual comparison between policies fine-tuned guided by PAVRM and our implicit process reward. These samples encompass a wider variety of semantic scenarios across both the 1.3B and 14B backbones. Consistent with the main text, the results underscore that GenRM-Flow provides a more reliable optimization signal, effectively mitigating the structural and temporal artifacts frequently observed in the baseline method.



Figure 5: Extended qualitative comparison of videos generated by policies fine-tuned via FlowAWR, utilizing either PAVRM or **GenRM-Flow** as the per-step reward signal. The top three rows display generations from models trained on the Wan2.1-T2V-1.3B backbone, while the bottom three rows exhibit results from models trained on the Wan2.1-T2V-14B backbone. Red bounding boxes highlight localized structural deformations, spatial inconsistencies, and temporal artifacts present in the PAVRM-guided outputs. Across both parameter scales and diverse semantic prompts, the proposed GenRM-Flow systematically yields generations with strictly maintained anatomical correctness and spatiotemporal continuity.

Table 7: Extended quantitative comparison including the sequence-level baseline (VideoAlign). **Bold** denotes the highest metric value per backbone.

Generative Model	Method	Inner Test Set				VBench [10]			
		MS ↑	DD↑	SC↑	HA↑	MS↑	DD↑	SC↑	HA↑
Wan2.1-T2V-1.3B	Pretrain	98.31	59.07	97.02	88.15	98.52	65.19	97.56	94.00
	VideoAlign [28]	98.58	60.68	97.22	89.96	98.73	67.52	97.64	95.32
	PAVRM [36]	98.56	63.59	97.18	89.73	98.71	67.83	97.52	94.68
	GenRM-Flow	**98.88**	**72.43**	**98.28**	**93.16**	**98.95**	**73.46**	**97.70**	**96.25**
Wan2.1-T2V-14B	Pretrain	98.12	61.18	97.18	89.32	98.30	65.46	97.52	95.40
	VideoAlign [28]	98.26	65.32	97.45	90.58	98.51	70.36	97.60	95.62
	PAVRM [36]	98.34	68.59	97.82	91.45	98.41	71.82	97.68	96.18
	GenRM-Flow	**98.76**	**77.48**	**98.45**	**95.08**	**99.03**	**74.87**	**97.73**	**97.73**

C Dataset Construction Details

This section provides extended details on the data collection and annotation protocols referenced in Section 4.

C.1 Dataset Statistics

Table 2 details the composition of the 48,757 raw videos generated prior to the annotation phase. We utilized 10 distinct text-to-video generation models. While the models exhibit varying native frame rates, resolutions, and output durations, all resulting videos were systematically normalized to a unified resolution of 832×480 at 16 fps with a 5 s duration prior to human evaluation.

C.2 Annotation Guidelines

Annotation Protocol. We employ 40 trained annotators to evaluate pairwise video quality. For each pair of videos generated from the same prompt, annotators select the preferred video based on overall motion quality. The annotation protocol is structured around three evaluation dimensions, summarized in Table 3.

Annotators evaluate each video pair holistically and select the video with higher overall motion quality. In cases where both videos are of comparable quality, annotators may label the pair as a tie; such pairs are excluded from the final dataset. To ensure annotation reliability, we apply majority voting among multiple annotators and filter out pairs with low inter-annotator agreement.

D Extended Baseline Comparisons in FlowAWR

In Section 4.2, the main text limits the evaluation of downstream generation quality to process-aware reward models (PAVRM and ours) to ensure structural symmetry in the FlowAWR feedback mechanism. For completeness, Table 7 provides the extended evaluation encompassing VideoAlign [28], which operates as a VLM-based reward baseline.

The empirical data indicates that our implicit process reward configuration (**Ours**) structurally outperforms VideoAlign across all evaluated dimensions on both parameter scales. Furthermore, the data reveals a scaling discrepancy in the PAVRM optimization signal: while PAVRM outperforms VideoAlign on the 14B backbone, it exhibits relative degradation compared to VideoAlign on the 1.3B backbone across three of the four VBench metrics (MS, SC, and HA).

E Prompts in Figures

The prompts used in Figure 1 are as follows:

1. A young girl with curly brown hair sits in a bright bedroom, wearing white over-ear headphones and a black camisole with white lace trim. She listens to music with her eyes closed, gently swaying and raising one hand to the rhythm. Behind her, a bed is visible along with a glass vase of flowers, and a small potted plant on the windowsill. Warm natural daylight streams in from a window on the left. The camera holds a static medium close-up shot at eye level. Soft indoor lighting, shallow depth of field.
2. A young man and a young woman sitting side by side at a black grand piano, performing a four-hand piano duet in a bright, modern living room. The man wears a dark navy sweater and has short black hair, his fingers moving precisely across the keyboard. The woman wears a beige knit cardigan with her hair pulled back in a ponytail, playing the piano with focused concentration. The camera holds a medium shot from a slightly elevated right-side angle, slowly pulling back to reveal both performers and the full length of the piano. Warm, natural indoor lighting.
3. Low angle shot, two skateboarders riding on an asphalt road. The primary focus is on the skateboarder in the right foreground, wearing a dark purple long-sleeve shirt, beige pants, and blue sneakers. He crouches deeply on his board to gather momentum, sharply pops the tail of the skateboard with his back foot, and leaps upwards to perform an ollie. The skateboard lifts off the ground and levels out perfectly in mid-air beneath his feet. He then lands smoothly back on the asphalt, bending his knees to absorb the impact. In the background to the left, a second skateboarder wearing a blue denim shirt, dark pants, and black sneakers continues riding steadily forward alongside him. Urban outdoor setting with a grassy curb in the background.
4. A bearded man in his thirties, wearing a plaid shirt and jeans, sits at a rustic wooden bar, surrounded by an array of beer taps and vintage brewery decor. He carefully lifts a frosty pint glass filled with amber beer, examining its color and clarity against the warm, ambient lighting. He takes a slow, appreciative sip, his eyes closing momentarily as he savors the complex flavors. The camera captures the subtle smile of satisfaction on his face, highlighting the rich foam on his upper lip. The background hum of soft chatter and clinking glasses adds to the cozy, inviting atmosphere of the pub.
5. A fluffy golden retriever running energetically towards the camera across a lush green grassy park field, carrying a bright orange frisbee in its mouth. The dog's fur, ears, and tail bounce dynamically with its running movement. The background features a spacious lawn with scattered trees under bright, natural daylight. Smooth tracking shot keeping the dog in the center of the frame, highly detailed, vibrant colors.
6. Two brown wild rabbits leap in sequence along a moss-covered riverbank in a lush green woodland setting. The first rabbit lands while the second follows mid-jump, their hind legs kicking up small splashes of water droplets from the damp ground. A calm, dark body of water occupies the foreground, reflecting the green vegetation. The background features dense, dark foliage under soft overcast daylight. The camera holds a low-angle side view at ground level, tracking the rabbits as they bound from left to right across the mossy bank. Natural lighting, shallow depth of field.
7. A Ghibli-inspired 3D animated video shows a red horned fantasy dragon with a yellow underbelly, flying forward towards the camera over a deep mountain valley. The dragon has curved horns and an open mouth in a roar, with large wings outstretched and flapping rhythmically. The landscape below is composed of steep green mountains and cliffs covered in small trees, with a vast green field in the center. The sky shows a warm sunset glow with pink, purple, and blue clouds scattered throughout. The camera is positioned directly in front of the dragon and tracks its motion forward, keeping the dragon centered in a dynamic tracking shot.
8. An animated, Ghibli-esque video sequence featuring a young Japanese woman with long black hair, wearing a blue floral print t-shirt and blue shorts. She stands at a large, open window

of a moving red and grey passenger train, looking back over her shoulder with a thoughtful expression as the train moves along rural tracks. Outside, a beautiful, changing landscape of golden fields, green hills, and distant mountains under a warm, cloudy evening sky unfolds. The camera maintains a stable, medium shot perspective focused on her and the view from the window as the train travels, with the fields and landscape passing by rapidly. The style is classic traditional anime, with soft colors and clear outlines.

The prompts used in Figure 4 are as follows:

1. A person is skydiving.
2. A person is rock climbing.
3. A person riding a majestic white horse, their attire blending seamlessly with the horse's coat, both animals moving gracefully across a rolling green pasture. The person has tousled brown hair and expressive blue eyes, their posture confident and relaxed as they hold the reins with steady hands. The horse's mane flows freely, catching the morning sunlight, adding a vibrant glow to the scene. In the background, lush trees and wildflowers dot the landscape, creating a serene and picturesque atmosphere. The person is wearing a simple yet elegant olive green tunic and sturdy leather boots, perfectly suited for the outdoors. The camera follows the duo, capturing moments of interaction between the rider and the horse, including subtle nods and playful glances. The scene transitions from a wide shot of the horse and rider to a medium shot focusing on the person's face, highlighting their joy and connection with nature. Cinematic lighting with soft shadows and warm tones enhances the emotional depth of the moment.
4. A cute and happy Corgi playing joyfully in a picturesque park during a beautiful sunset, captured in super slow motion. The Corgi, with its fluffy white fur and expressive brown eyes, runs swiftly across the green grass, wagging its tail enthusiastically. Its playful demeanor is evident as it jumps over small obstacles, rolling in the soft grass, and occasionally sniffing at new scents. The sun sets behind the park, casting a warm golden glow over everything, creating a serene and tranquil atmosphere. The sky transitions from vibrant shades of orange and pink to deep purples and blues, adding a magical touch to the scene. The camera captures every moment, from the Corgi's joyful expressions to the gentle rustling of leaves, all in slow motion, allowing viewers to savor each second of this delightful moment.
5. CG game concept digital art, a muscular man stretching one of his long legs. He is wearing a tight-fitting black t-shirt and faded blue jeans. His muscles ripple as he stretches, revealing his toned physique. The lighting is soft and natural, casting subtle shadows on his body. He is standing on a rugged stone path, surrounded by tall grass and wildflowers. The background features a dense forest with ancient trees and vibrant greenery. Low-angle view, focusing on the full stretch of his leg. Close-up shot emphasizing his facial expression, showing determination and relaxation.
6. A whimsical scene captured in a vibrant watercolor style, showcasing a majestic bird perched gracefully atop a slender tree branch, its feathers shimmering in iridescent hues of blue and green. The bird has piercing yellow eyes and a curved beak, its wings slightly spread for balance. Nearby, a curious cat peers over the edge of a nearby roof, its fur a mix of soft gray and black, eyes wide with interest. The cat has tufted ears and a fluffy tail, standing on its hind legs as it investigates the bird. The background is a lush garden filled with blooming flowers and tall grasses, with a gentle sunbeam casting dappled shadows. Soft pastel colors and gentle brushstrokes create a serene and peaceful atmosphere. Medium shot, side-by-side composition.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction precisely reflect the contribution and scope of this paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have created a separate "Limitations and future work" section in our paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide a complete proof for each theoretical result in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We present the detailed experimental settings in the Section 4 and we will release the code and checkpoint soon.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the code and checkpoint soon.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the training and test details in the Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We use the statistical significance in the evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We point out the specific compute resources in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper obeys the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no concern about broader impacts involved in this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There is no concern about safeguards involved in this work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit all works cited including dataset, code and model in the references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets (e.g., code and checkpoint) will be released soon.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology does not involve the use of large language models. LLMs were only used for language polishing and writing assistance, which does not affect the scientific contributions of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.