FLARE: IMPROVING LONG-HORIZON FIDELITY IN AUTOREGRESSIVE VIDEO GENERATION VIA FACTORIZED LATENT REWARD FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video diffusion with streaming training enables real-time long-form generation, but error accumulation across chunks makes long-horizon consistency a fundamental challenge. Existing streaming distillation objectives keep generated chunks on the natural-video manifold, but provide limited guidance for resolving quality and alignment ambiguities during autoregressive rollout. As a result, long rollouts often suffer from semantic drift, visual quality decay, and motion degradation. We introduce FLARE, which closes this gap through a latent preference reward that scores video latents directly via a DiT backbone, equipped with three learnable reward queries that attend to backbone features and route to independent heads, allowing a single video pair to express conflicting preferences across aspects. The reward is trained in two stages: a generic Bradley-Terry preference SFT, followed by a consensus-guided latent preference adaptation stage in which multiple pixel-space evaluators serve as weak head-wise annotators and only cross-evaluator-agreed comparisons are retained. The trained reward is then integrated as an online, differentiable per-chunk signal alongside DMD distillation, providing aspect-specific preference guidance during streaming optimization while retaining distribution-level supervision from DMD. Experiments on long-video generation show that FLARE improves generation fidelity over strong autoregressive baselines, and ablations verify the effectiveness of the latent reward design.

1 INTRODUCTION

Video generation has achieved remarkable quality in short-clip synthesis (HaCohen et al., 2026; Wan et al., 2025; Kong et al., 2024), driven by advances in Diffusion Transformer (DiT)-based models. Extending these capabilities to long-form, interactive generation—where identity, composition, and prompt semantics must remain coherent—remains an open challenge. Autoregressive (AR) generation (Yin et al., 2025) addresses this by producing latent chunks sequentially, each conditioned on a KV cache encoding the full generation history. Streaming training (Huang et al., 2025; Cui et al., 2026; Yang et al., 2026a; Zhu et al., 2026) mirrors this inference procedure by optimizing the generator on its own rollout state, eliminating the train–test mismatch that plagues short-horizon models deployed at long context. This formulation enables real-time, causally conditioned video generation, but it also exposes a central training challenge: each generated chunk affects all future chunks, so small local errors can accumulate into long-range semantic and visual degradation.

Despite training on self-generated rollouts, streaming AR models still exhibit systematic degradation as the generation horizon grows (Fig. 1). These failures typically appear in three forms: ***semantic drift***, where subjects duplicate, identities change, attributes mutate, or the scene composition gradually departs from the prompt; ***visual quality decay***, including blur, color banding, texture collapse, and structural artifacts that become more prominent at later chunks; and ***motion degeneration***, with object teleportation or unnatural speed changes. Although Distribution Matching Distillation (DMD) (Yin et al., 2024) anchors the generator to the conditional video distribution, it does not explicitly distinguish which plausible continuation better preserves semantic alignment, visual quality, and dynamic quality over time. Moreover, these failure modes need not occur simultaneously: a chunk may preserve visual quality while semantically drifting, or preserve semantic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

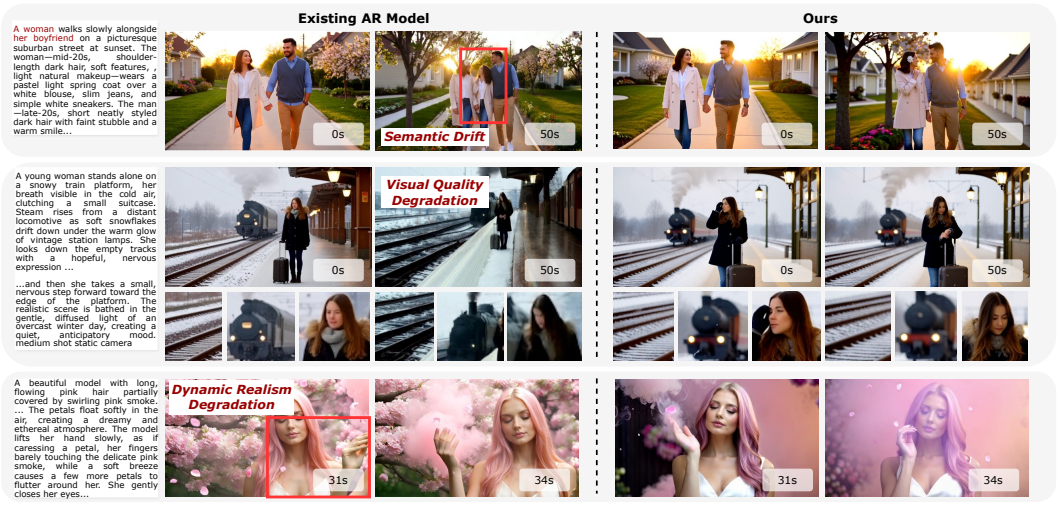

Figure 1: Long-horizon degradation under streaming rollout. Each row shows chunk-index progression for one failure category: semantic drift, visual quality degradation, and dynamic realism degradation.

alignment while exhibiting unstable motion. Long-horizon generation therefore requires not only distributional regularization, but also directional and aspect-specific preference supervision.

Existing reward mechanisms are poorly suited to this setting. Some reinforcement-learning methods (Liu et al., 2025a; Xue et al., 2025) can optimize black-box video rewards, but typically estimate gradients from sampled trajectories rather than providing direct supervision at each generated chunk. Differentiable pixel-space rewards (Liu et al., 2025b; Wang et al., 2026) offer denser feedback, yet require back-propagation through VAE decoding and large visual encoders, resulting in prohibitive activation memory during long streaming rollouts. In addition, collapsing semantic alignment, visual quality, and dynamic quality into a single scalar judgment may obscure conflicting preferences across these aspects. These limitations motivate a reward formulation that is compatible with long streaming rollouts while preserving aspect-specific preference signals.

To this end, we introduce FLARE, which combines a factorized latent-space reward with streaming DMD optimization. FLARE learns a differentiable preference reward directly on video VAE latents using a pretrained video DiT backbone. Three learnable reward queries attend to shared spatiotemporal features and route them to independent heads for Semantic Alignment, Visual Quality, and Dynamic Quality, allowing a single video pair to express different preference directions across aspects. The reward model is trained in two stages: generic Bradley–Terry preference supervised fine-tuning, followed by domain-specific preference distillation in which multiple pixel-space evaluators serve as weak head-wise annotators and only cross-evaluator-agreed comparisons are retained. During generator optimization, the trained reward model is frozen and applied directly to the generator's streaming rollout, providing an online differentiable per-chunk loss alongside DMD. In this way, DMD preserves the underlying video distribution, while the latent reward steers generation toward more faithful continuations within it.

The contributions of this work are summarized as follows:

- We introduce a factorized latent reward model for autoregressive video streaming, which predicts semantic alignment, visual quality, and dynamic quality directly from video latents and provides differentiable per-chunk feedback without pixel-space decoding during generator optimization.
- We introduce a consensus-guided head-wise distillation scheme that trains this reward from cross-evaluator agreement of multiple pixel-space critics.
- We present FLARE, which integrates the reward as a differentiable per-chunk loss alongside DMD streaming distillation.
- Experiments show that our proposed method consistently improves long-horizon fidelity across interactive and single-prompt video generation settings.

2 Related Work

Autoregressive Video Generation. Autoregressive (AR) video diffusion (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026) reformulates video synthesis as chunk-by-chunk generation conditioned on a persistent KV cache, replacing fixed-window training with an open-ended streaming process. Self-Forcing (Huang et al., 2025) closes the train–test gap inherent in this formulation by optimizing the generator on its own rollout state via Distribution Matching Distillation (Yin et al., 2024), and a line of successors (Cui et al., 2026; Yang et al., 2026a; Liu et al., 2025c; Yang et al., 2026b; Lu et al., 2025) extends the paradigm toward minute-scale and real-time interactive generation. Recent work focuses on adaptive memory mechanisms (Mao et al., 2026; Ji et al., 2025) and positional-encoding designs (Yesiltepe et al., 2026) to sustain coherence as generation unfolds beyond its training horizon.

Reinforcement Learning for Video Generation. Reward-based post-training has emerged as a complementary supervision paradigm for video generation. Policy-gradient methods adapted to flow-matching and diffusion samplers (Xue et al., 2025; Liu et al., 2025a; Zheng et al., 2025) optimize generators against scalar rewards. Differentiable reward backpropagation (Prabhudesai et al., 2024) chains analytic per-sample gradients from the reward through the sampler back to the generator, providing dense, low-variance direction at the cost of requiring a differentiable reward and end-to-end activation memory. Recent work extends this paradigm to distilled AR streaming generators (Yang et al., 2026b; He et al., 2026; Zhang et al., 2026; Lu et al., 2026; 2025), innovating on the RL machinery—group sampling, chunk-level forking, forward-process formulations—and often summing multiple pretrained pixel evaluators to form a multi-aspect reward. These methods estimate reward gradients from trajectory-level samples and accept any black-box reward. The reward signal is typically supplied by VLM-based pixel-space evaluators (Liu et al., 2025b; Wang et al., 2026; Cao et al., 2026), each carrying its own scale and rubric. The differentiable line, however, must retain activations through VAE decoding and VLM re-encoding to keep the gradient chain alive, capping the rollout length that fits in memory. Our work is complementary to rewarded streaming optimization methods that incorporate pretrained pixel-space rewards into autoregressive video training. Rather than directly optimizing such evaluators, FLARE distills their aspect-specific preference directions into a factorized latent reward model and applies the resulting differentiable reward directly to streaming latent chunks. This design targets efficient per-chunk preference feedback while avoiding VAE decoding and visual re-encoding in the reward gradient path.

3 Method

3.1 Preliminary

Streaming Autoregressive Video Diffusion. Autoregressive video diffusion models generate videos sequentially in latent space. Given a prompt c and latent video $\mathbf{x}_{1:T} = \{x_1, \ldots, x_T\}$, the causal generation process is factorized as

$$p_\theta(\mathbf{x}_{1:T} \mid c) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}, c). \tag{1}$$

In practice, the model predicts a short chunk of future latents conditioned on previously generated frames. Mainstreaming methods (Huang et al., 2025; Yang et al., 2026a) implement this process with a causal Wan (Wan et al., 2025) generator and a persistent KV cache. At step t, the streaming state is $s_t = \{\mathcal{C}_t, t, c\}$, where $\mathcal{C}_t$ stores historical KV states. The generator samples noise for the next chunk, denoises it with causal attention over the cache, appends the clean prediction to the stream, and updates the cache for subsequent generation. Training follows the same self-rollout process to reduce train-test mismatch. Each update forms a local training window

$$\mathbf{z}_t = [\mathrm{sg}(\mathbf{z}^{\mathrm{hist}}_{t-o:t-1}), \mathbf{z}^{\mathrm{new}}_{t:t+n-1}], \tag{2}$$

where o and n denote the numbers of overlap and newly generated latent frames, respectively, and $\mathrm{sg}(\cdot)$ is the stop-gradient operator. The overlap region provides temporal context, whereas gradients are retained only for the newly generated region.

Following prior streaming autoregressive video generation method, we employ Distribution Matching Distillation (DMD) (Yin et al., 2024) to align the few-step generator with the pretrained video

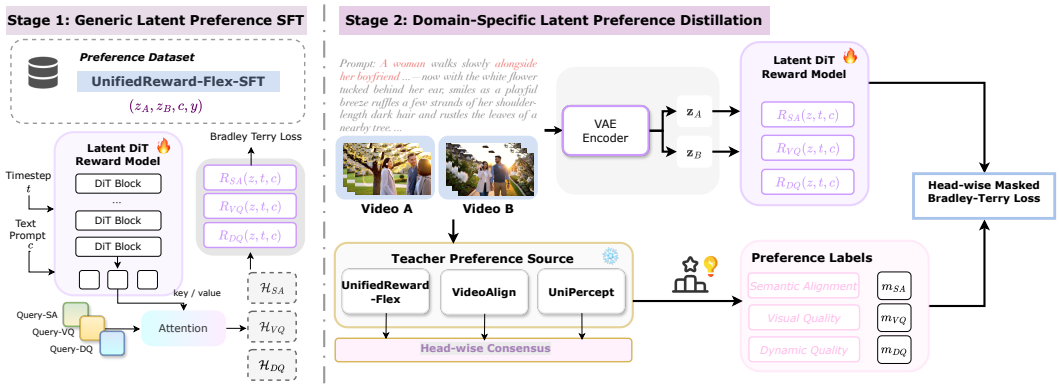

Figure 2: Overview of factorized latent reward training. The latent reward model is first pretrained on generic video preference pairs and then adapted to the autoregressive long-video domain using head-wise consensus supervision from multiple pixel-space evaluators. Aspect-specific queries and heads predict semantic alignment, visual quality, and dynamic quality from shared latent DiT features.

diffusion model. Let $\mathbf{z}_\theta = G_\theta(\boldsymbol{\epsilon}, c)$ denote the clean latent generated from Gaussian noise $\boldsymbol{\epsilon}$. We sample a diffusion timestep τ and re-noise the generated latent as $\widetilde{\mathbf{z}}_\tau = \alpha_\tau \mathbf{z}_\theta + \sigma_\tau \boldsymbol{\epsilon}_\tau$. DMD minimizes the reverse KL divergence between the generator distribution and the target data distribution. Its gradient with respect to the generator parameters can be approximated by the difference between the score functions of the generated and real distributions:

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}_\tau}\left[w(\tau)\left(s_{\mathrm{fake}}(\widetilde{\mathbf{z}}_\tau, c, \tau) - s_{\mathrm{real}}(\widetilde{\mathbf{z}}_\tau, c, \tau)\right) \frac{\partial \mathbf{z}_\theta}{\partial \theta} \right], \tag{3}$$

where s_{real} is estimated by a frozen pretrained diffusion model, s_{fake} is estimated by an online score model trained on generated samples, and $w(\tau)$ is a timestep-dependent weighting function.

3.2 Overview

We propose FLARE, a framework that improves long-horizon autoregressive video generation through Factorized Latent Reward Feedback. FLARE first learns a differentiable reward model that evaluates video latents along semantic, visual, and dynamic dimensions. The learned reward is then applied to the generator's own streaming rollout as chunk-level supervision alongside Distribution Matching Distillation (DMD).

3.3 Factorized Latent Reward Modeling

We learn a differentiable reward model that operates directly on video VAE latents. During generator optimization, the latent reward replaces expensive pixel-space evaluators, avoiding repeated VAE decoding and visual re-encoding. The reward model evaluates each video along three aspects: Semantic Alignment (SA), Visual Quality (VQ), and Dynamic Quality (DQ).

3.3.1 Architecture

Given a video latent $\mathbf{z}_t$ at diffusion timestep t and a text condition c, we use a pretrained video DiT backbone to extract spatiotemporal representations:

$$\mathbf{H} = f_\phi\left(\mathbf{z}_t, t, c\right), \qquad \mathbf{H} \in \mathbb{R}^{N \times D}, \tag{4}$$

where ϕ denotes the reward-model parameters. We define the set of reward aspects as $\mathcal{A} = \{\mathrm{SA}, \mathrm{VQ}, \mathrm{DQ}\}$. For each aspect $a \in \mathcal{A}$, we introduce a learnable reward query $\mathbf{q}_a$. The query attends to the shared DiT features through query-attention pooling:

$$\mathbf{u}_a = \mathrm{Attn}\left(\mathbf{q}_a, \mathbf{H}\mathbf{W}_K, \mathbf{H}\mathbf{W}_V\right), \qquad a \in \mathcal{A}, \tag{5}$$

where $\mathbf{W}_K$ and $\mathbf{W}_V$ are shared key and value projections. Each aspect-specific representation is mapped to an independent scalar reward:

$$R_\phi^a(\mathbf{z}_t, t, c) = g_\phi^a(\mathbf{u}_a), \qquad a \in \mathcal{A}. \tag{6}$$

The shared DiT backbone captures general spatiotemporal information, while the aspect-specific queries and heads preserve distinct preference signals for semantic alignment, visual quality, and dynamic quality. To expose the reward model to different latent diffusion states, we perturb the clean video latent $\mathbf{z}$ according to $u \sim \mathcal{N}(-4, 1^2)$, $t = \mathrm{sigmoid}(u)$, $\mathbf{z}_t = \alpha_t \mathbf{z} + \sigma_t \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Here, u has mean -4 and standard deviation 1, inducing a Logit-Normal distribution over t that emphasizes low-noise diffusion states. This perturbation acts as latent-space augmentation and encourages more stable preference predictions.

3.3.2 Generic Latent Preference SFT

We initialize the latent reward model from a pretrained Wan2.1-T2V-1.3B DiT and perform full-parameter preference supervised fine-tuning on 45K video preference pairs from UnifiedReward-Flex-SFT (Wang et al., 2026). During this stage, the DiT backbone, query-attention pooling module, and three aspect-specific reward heads are jointly optimized. Each training sample contains a pair of videos and category-level preferences over three dimensions: semantic alignment (SA), visual quality (VQ), and Dynamic Quality (DQ). For aspect a, let $\mathbf{z}_{w,t}^a$ and $\mathbf{z}_{l,t}^a$ denote the preferred and dispreferred video latents at timestep t, respectively. We optimize a head-wise Bradley–Terry objective:

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}\left[\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \log\sigma\left(R_\phi^a\left(\mathbf{z}_{w,t}^a, t, c\right) - R_\phi^a\left(\mathbf{z}_{l,t}^a, t, c\right)\right)\right]. \tag{7}$$

This stage provides a general latent preference initialization before adaptation to the target autoregressive video domain.

3.3.3 Consensus-Guided Latent Preference Adaptation

Generic preference data do not fully capture the degradation patterns of long-horizon autoregressive generation. We therefore construct target-domain video pairs and adapt the reward model using preference signals from multiple pixel-space evaluators.

Target-domain pair construction. We generate 60-second videos using representative autoregressive long-video models, including LongLive (Yang et al., 2026a), MemFlow (Ji et al., 2025), and Anchor Forcing (Yang et al., 2026b). Each video consists of six consecutive 10-second prompt segments. We divide each segment into two 5-second clips and pair them with same-prompt short videos generated by Causal Forcing(Zhu et al., 2026) and Self-Forcing (Huang et al., 2025). We use VideoAlign (Liu et al., 2025b), UnifiedReward-Flex (Wang et al., 2026), and UniPercept (Cao et al., 2026) as complementary preference sources. Visual-quality preferences are determined using the corresponding scores from UnifiedReward-Flex and UniPercept. Dynamic-quality preferences are determined using UnifiedReward-Flex and VideoAlign. Semantic-alignment preferences are determined using the alignment scores from UnifiedReward-Flex and VideoAlign.

Head-wise consensus. Because different evaluators use different scoring scales and rubrics, we convert their outputs into pairwise preference directions instead of directly regressing their scalar scores. For evaluator j and aspect a, the pairwise score difference is

$$\Delta_j^a = S_j^a(\mathbf{v}_A, c) - S_j^a(\mathbf{v}_B, c), \tag{8}$$

where $\mathbf{v}_A$ and $\mathbf{v}_B$ denote the two decoded videos. Evaluator j abstains when $\left|\Delta_j^a\right| < \delta_j^a$, where δ_j^a is an evaluator-specific confidence margin. For each aspect, the consensus procedure produces a preference label y_a and a reliability mask $m_a \in \{0, 1\}$. We set $m_a = 1$ when the non-abstaining evaluators associated with aspect a agree on the preference direction. When $m_a = 1$, $y_a = +1$ indicates that video A is preferred over video B, while $y_a = -1$ indicates the opposite. Given the corresponding latent pair $(\mathbf{z}_{A,t}, \mathbf{z}_{B,t})$, we optimize the masked head-wise Bradley–Terry loss:

$$\mathcal{L}_{\mathrm{adapt}} = \frac{\sum_{a\in\mathcal{A}} m_a\left[-\log\sigma\left(y_a\left[R_\phi^a(\mathbf{z}_{A,t}, t, c) - R_\phi^a(\mathbf{z}_{B,t}, t, c)\right]\right)\right]}{\sum_{a\in\mathcal{A}} m_a + \epsilon}. \tag{9}$$

Each reward head is therefore supervised independently, allowing different aspects of the same video pair to have different preference directions.

3.4 Streaming Alignment with Latent Reward Feedback

After reward-model training, we freeze ϕ and use the latent reward as a differentiable critic during streaming generator optimization.

3.4.1 Boundary-Aware Reward Window

At streaming step k, the reward model evaluates a local latent window consisting of one overlap frame from the previous chunk and the newly generated chunk:

$$\mathbf{z}_k^{\mathrm{RM}} = \left[\mathrm{sg}\left(\mathbf{z}_{k-1}^{\mathrm{ovlp}}\right), \mathbf{z}_k^{\mathrm{new}}\right]. \tag{10}$$

The overlap frame serves as a temporal anchor for evaluating the new chunk relative to the preceding rollout. Since it is detached, reward gradients are propagated only through the newly generated frames. We perturb the reward window to a sampled diffusion timestep t: $\mathbf{z}_{k,t}^{\mathrm{RM}} = \alpha_t \mathbf{z}_k^{\mathrm{RM}} + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where t follows the same $\mathrm{LogitNormal}(-4, 1)$ distribution used during reward-model training. For each aspect, the frozen reward model predicts

$$r_k^a = R_\phi^a\left(\mathbf{z}_{k,t}^{\mathrm{RM}}, t, c_k\right), \qquad a \in \mathcal{A}. \tag{11}$$

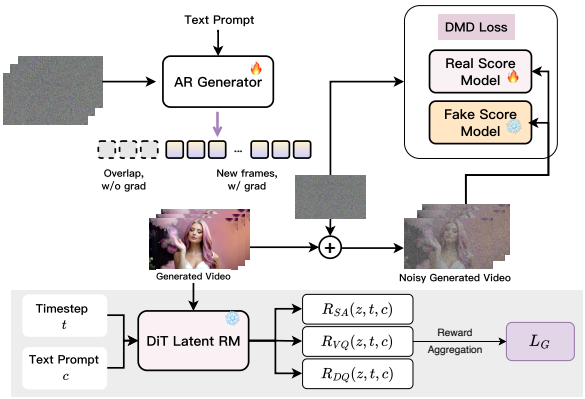

Figure 3: Streaming alignment with factorized latent reward feedback. The frozen reward model evaluates one detached overlap frame together with the newly generated chunk and provides differentiable aspect-wise feedback. The reward loss is combined with DMD to optimize the autoregressive generator.

3.4.2 Generator Objective

Because the three reward heads may have different output scales, we normalize their predictions using statistics computed on a held-out validation set: $\widehat{r}_k^a = \frac{r_k^a - \mu_a}{\sigma_a + \epsilon}$, where μ_a and σ_a denote the validation mean and standard deviation for aspect a. The normalized aspect rewards are aggregated as $R_k = \sum_{a \in \mathcal{A}} \lambda_a \widehat{r}_k^a$, where λ_a controls the contribution of each aspect. The reward loss for generator optimization is $\mathcal{L}_{\mathrm{reward}} = -R_k$. Although the reward-model parameters are frozen, the reward remains differentiable with respect to the generated latent chunk:

$$\nabla_\theta \mathcal{L}_{\mathrm{reward}} = -\frac{\partial R_k}{\partial \mathbf{z}_{k,t}^{\mathrm{RM}}} \frac{\partial \mathbf{z}_{k,t}^{\mathrm{RM}}}{\partial \mathbf{z}_k^{\mathrm{new}}} \frac{\partial \mathbf{z}_k^{\mathrm{new}}}{\partial \theta}. \tag{12}$$

The final generator objective combines distribution matching and latent reward feedback:

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{DMD}} + \lambda_{\mathrm{reward}} \mathcal{L}_{\mathrm{reward}}. \tag{13}$$

We use equal weights for the three reward heads. DMD provides distribution-level supervision, while the factorized latent reward provides aspect-specific preference feedback throughout the autoregressive rollout.

4 Experiments

4.1 Experiment Setup

Implementation details. We use Wan2.1-T2V-1.3B (Wan et al., 2025) as the base autoregressive video generator and conduct ODE initialization following Self-Forcing (Huang et al., 2025). The



Figure 4: Qualitative comparison results for interactive-prompt generation.

Table 1: Quantitative results on 60-second interactive long-horizon video generation.

Method	Sub.↑	Back.↑	Motion↑	Imaging↑	CLIP Score↑					
					0–10 s	10–20 s	20–30 s	30–40 s	40–50 s	50–60 s
Self Forcing	96.50	95.73	97.22	65.14	21.15	20.52	19.73	19.15	19.33	19.66
LongLive	97.77	96.13	99.09	68.52	**21.99**	21.34	21.30	20.13	20.45	19.64
Infinity-RoPE	97.93	96.27	99.08	68.79	**21.99**	21.54	21.45	20.33	21.22	19.18
MemFlow	97.04	95.21	98.79	65.73	21.87	21.22	20.36	20.03	20.44	19.25
Anchor Forcing	96.49	96.01	97.95	68.35	21.48	20.46	19.89	19.90	19.79	17.59
Ours	**98.02**	**96.57**	**99.17**	**69.13**	21.88	**21.60**	**21.56**	**20.50**	**21.41**	**20.01**

generator is then optimized with DMD under the LongLive (Yang et al., 2026a) streaming protocol. The training prompts are derived from the VidProM (Wang & Yang, 2024) dataset following LongLive (Yang et al., 2026a). Videos are generated at 832×480 resolution. For the latent reward model, we adopt the same DiT backbone family as the generator. In the Consensus-Guided Latent Preference Adaptation stage, we fine-tune the reward model with LoRA of rank 16 while training the query pooling modules and reward heads. For generator optimization, the trained reward model is frozen and used only as a differentiable latent-space critic. The three reward scores are normalized with held-out validation statistics and combined as the reward loss. We use the same prompts, resolution, rollout length, and evaluation protocol for all autoregressive baselines.

4.2 Comparison Results

4.2.1 Interactive Long-Horizon Video Generation.

To evaluate interactive long-horizon video generation, we build 100 narrative sequences based on the LongLive protocol (Yang et al., 2026a). Each sequence is divided into six ordered 10-second prompt segments, and the model is required to generate a 60-second video for each sequence. Tab. 1 presents the quantitative comparison against Self Forcing (Huang et al., 2025), Infinity-RoPE (Yesiltepe et al., 2026), LongLive (Yang et al., 2026a), MemFlow (Ji et al., 2025), and Anchor Forcing (Yang et al., 2026b).



Figure 5: Qualitative ablation results for single-prompt 60s-video generation.

Table 2: Quantitative comparison of video generation models (5s).

Model	Params.	Res.	Evaluation scores ↑		
			Total	Quality	Semantic
Self Forcing	1.3B	832×480	83.42	83.97	81.22
LongLive	1.3B	832×480	83.27	83.57	82.06
Rolling Forcing	1.3B	832×480	82.54	83.34	79.33
MemFlow	1.3B	832×480	83.64	84.44	80.46
Reward Forcing	1.3B	832×480	83.86	84.58	80.99
Anchor Forcing	1.3B	832×480	83.59	84.40	80.35
Ours	1.3B	832×480	**84.50**	**84.98**	**82.56**

Table 3: Comparison results for 30s single-prompt long-video generation on VBench-Long.

Model	Total Score↑	Quality Score↑	Semantic Score↑
Self-Forcing	83.18	84.07	79.62
MemFlow	83.19	84.03	79.86
LongLive	82.63	83.25	80.15
Reward Forcing	83.74	84.71	79.85
Anchor Forcing	84.44	85.40	80.64
Ours	**84.87**	**85.82**	**81.05**

As shown in Tab. 1, FLARE achieves the highest subject consistency, background consistency, motion smoothness, and imaging quality scores under continuously changing prompts. FLARE achieves the highest score in each of the subsequent five segments, indicating that its advantage becomes more pronounced as the autoregressive rollout progresses. The examples in Fig. 4 show more stable subject appearance and scene composition over the 60-second rollout.

4.2.2 Short Video Generation.

We also evaluate the proposed method under the conventional 5s-video generation setting to verify its general generation capability. Following the standard VBench protocol, we use the official prompt set. As summarized in Tab. 2, our method achieves the best overall score on 5-second video generation, showing that the model retains strong short-clip generation capability

Table 4: Comparison results for 60s single-prompt long-video generation.

Model	Subject Consis.	Background Consis.	Motion Smooth.	Aesthetic Quality	Imaging Quality
Self-Forcing	95.17	95.39	97.57	59.17	66.85
MemFlow	97.27	96.55	98.49	60.12	67.80
LongLive	**97.82**	96.90	98.76	61.69	68.95
Reward Forcing	97.16	96.55	98.24	60.29	69.28
Anchor Forcing	96.54	96.38	97.43	60.28	68.38
Ours	97.73	**97.17**	**98.86**	**62.76**	**70.33**

Table 5: Quantitative comparison of reward models on GenAI-Bench-Video and our human-annotated target-domain test set of 400 video pairs.

Method	GenAI-Bench	Test Set		
		VQ	DQ	SA
VideoAlign	68.85	66.50	65.00	59.25
UnifiedReward-Flex	**78.77**	62.00	69.00	67.25
Ours	76.50	**67.75**	**74.00**	**74.50**
Ours w/o CLPA	75.50	62.25	73.50	66.50
Ours w/o SFT + w/o CLPA	51.71	48.00	39.75	50.75

4.2.3 Long Video Generation.

We further evaluate single-prompt long-video generation to verify whether the proposed latent reward feedback improves long-horizon fidelity beyond the interactive-prompt setting. For 30-second generation, we follow the VBench-Long (Huang et al., 2024) protocol and report the overall, qual-

ity, and semantic scores in Tab. 3. Compared with other autoregressive baselines, our method achieves better long-video performance. For 60-second generation, we use the first 128 prompts from MovieGen (Polyak et al., 2024) following CausVid (Yin et al., 2025) and evaluate multiple long-video dimensions. As shown in Tab. 4, our method improves the most relevant long-horizon consistency metrics while maintaining visual quality.

Table 6: Ablation study of the latent reward model (Preference Accuracy (%)).

Method	VQ	DQ	TA
w/o noise + w/o CLPA	61.50	73.50	74.00
w/o noise	61.25	67.25	66.50
w/ noise + w/o CLPA	62.25	73.50	66.50
Ours	**67.75**	**74.00**	**74.50**

Table 7: Ablation study of factorized reward modeling.

Method	Quality↑	Semantic↑
DMD only	84.16	79.98
Single-head RM	84.87	80.65
Factorized RM **(Ours)**	**85.82**	**81.05**

4.3 Ablation Studies

4.3.1 Reward Model Comparison

We evaluate the proposed latent reward model on GenAI-Bench-Video (Jiang et al., 2024) and an internal human-annotated test set. The internal test set contains 400 pairwise video samples generated from our target long-horizon autoregressive setting, with human preference labels collected independently for visual quality, dynamic quality, and text alignment. For GenAI-Bench-Video, samples annotated as ties are excluded, and the final decision of our model is obtained by majority voting over the three aspect heads. As shown in Tab. 5, our latent reward model remains competitive with strong pixel-space reward models on general preference evaluation, while achieving the best performance on the human-annotated internal test set. The ablation results further confirm the importance of the two-stage training strategy. Removing Consensus-Guided Latent Preference Adaptation (CLPA) reduces the accuracy on the human-annotated test set, especially for visual quality and text alignment. Without both generic preference SFT and CLPA, the model performs close to random across all aspects, indicating that preference supervision is essential for turning latent DiT representations into reliable reward signals.

4.3.2 Effect of the Latent Reward in Streaming Training

Compared with the model trained without latent reward feedback, the examples in Fig. 5 show improved visual details and more stable subject appearance. Together with the quantitative ablation in Tab. 7, these results support the benefit of latent reward feedback for long-video generation.

4.3.3 Effect of the Component of RM

We further ablate the training design of the latent reward model on the internal human-annotated test set. As shown in Tab. 6 and 7, combining noisy latent training with CLPA yields the strongest overall reward-model accuracy. In downstream generation, the factorized reward further outperforms the single-head variant, supporting the effectiveness of maintaining separate aspect-specific reward heads.

5 Conclusion

In this work, we presented FLARE, a long-horizon alignment framework for autoregressive video generation built around a latent preference reward. FLARE learns an aspect-aware reward model directly in latent space and integrates it into streaming DMD training as an online differentiable signal. This design provides directional supervision for mitigating semantic drift, visual quality decay, and motion degradation, while avoiding expensive pixel-space reward inference during generator optimization. Experiments on interactive and single-prompt long-video generation show consistent improvements over strong autoregressive baselines, and ablations validate the effectiveness of the latent reward design and the two-stage preference training strategy.

AI USE STATEMENT

In this work, we used generative AI tools to assist in constructing interactive multi-segment prompt sequences for long-horizon video generation. We additionally used generative AI tools for language editing. We did not use generative AI tools to generate research ideas, implement the proposed method, or produce experimental results. All AI-assisted content was reviewed by the authors, who take responsibility for the final content of this work.

REFERENCES

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, Xiangyu Chen, Wenlong Zhang, Bin Fu, Yu Qiao, and Yihao Liu. Unipercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. In *ICML*, 2026.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Dailan He, Guanlin Feng, Xingtong Ge, Yi Zhang, Bingqi Ma, Guanglu Song, Yu Liu, and Hongsheng Li. Ar-copo: Align autoregressive video generation with contrastive policy optimization. *arXiv preprint arXiv:2603.17461*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *NIPS*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

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.

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. In *NIPS*, 2025a.

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. In *NIPS*, 2025b.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025c.

Yanzuo Lu, Ronglai Zuo, and Jiankang Deng. Raven: Real-time autoregressive video extrapolation with consistency-model grpo. *arXiv preprint arXiv:2605.15190*, 2026.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Xiaofeng Mao, Shaohao Rui, Kaining Ying, Bo Zheng, Chuanhao Li, Mingmin Chi, and Kaipeng Zhang. Packforcing: Short video training suffices for long video sampling and long context inference. *arXiv preprint arXiv:2603.25730*, 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *NIPS*, 37:65618–65642, 2024.

Yibin Wang, Yuhang Zang, Feng Han, Jiazi Bu, Yujie Zhou, Cheng Jin, and Jiaqi Wang. Unified personalized reward model for vision generation. *arXiv preprint arXiv:2602.02380*, 2026.

Xiaomi. Mimo-v2.5. `https://huggingface.co/collections/XiaomiMiMo/mimo-v25`, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, and Song Hanand Yukang Chen. Longlive: Real-time interactive long video generation. In *ICLR*, 2026a.

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region rope for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026b.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

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

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A OVERVIEW

This is supplementary material for *FLARE: Improving Long-Horizon Fidelity in Autoregressive Video Generation via Factorized Latent Reward Feedback*.

We provide additional details and results, including the interactive prompt generation pipeline, target-domain preference pair construction, evaluator consensus statistics, human annotation protocol, computational scalability, and additional chunk-wise analyses.

A.1 INTERNAL HUMAN PREFERENCE TEST SET

We construct a held-out internal test set to evaluate the latent reward model in our target autoregressive video generation setting. The set contains 400 prompt-conditioned video pairs. As shown in Fig. A.3, each pair is displayed side by side together with the text prompt, and annotators independently select the preferred video along three aspects: text-video alignment, dynamic quality, and visual quality. Each pair is labeled by three annotators, and the order of Video A and Video B is randomized. Final labels are obtained by majority vote for each aspect.

Table A.1: Computational scalability of differentiable reward supervision.

Reward Type	Decode Required	Backprop Through VAE	Max Reward Window
VideoAlign	✓	✓	6
UnifiedReward-Flex	✓	✓	OOM
Ours	✗	✗	21

A.2 COMPUTATIONAL SCALABILITY

We compare the scalability of different differentiable reward supervision schemes in Tab. A.1. Pixel-space reward models such as VideoAlign (Liu et al., 2025b) and UnifiedReward-Flex (Wang et al., 2026) require decoding generated latents into videos and back-propagating gradients through the VAE decoder and the reward network. This substantially increases activation memory during streaming training and limits the reward window that can be optimized jointly. In contrast, our latent reward model operates directly on video latents and therefore removes the need for pixel-space decoding during reward optimization. As a result, it supports a longer reward window under the same training setting, enabling denser long-horizon feedback for autoregressive video generation.

A.3 ADDITIONAL EXPERIMENTS

Chunk-wise long-horizon analysis. The left panels of Fig. A.1 report metrics at different rollout chunks. Compared with autoregressive baselines, FLARE exhibits slower degradation in later chunks, especially for text alignment, visual quality, and motion smoothness, suggesting improved long-horizon fidelity.

Reward score evolution. The right panel reports the aggregated latent reward score during generator optimization. The score is computed by summing the three aspect-specific reward heads after normalization with fixed validation-set statistics. It increases steadily and grows more moderately in later iterations, indicating diminishing reward gains from further updates

Table A.2: Head-wise evaluator consensus on candidate pairs. Agreement is measured as the percentage of candidate pairs for which the non-abstaining evaluators assign the same preference direction.

Aspect	Consensus Agreement
Semantic alignment	78.4%
Dynamic quality	72.6%
Visual quality	81.7%

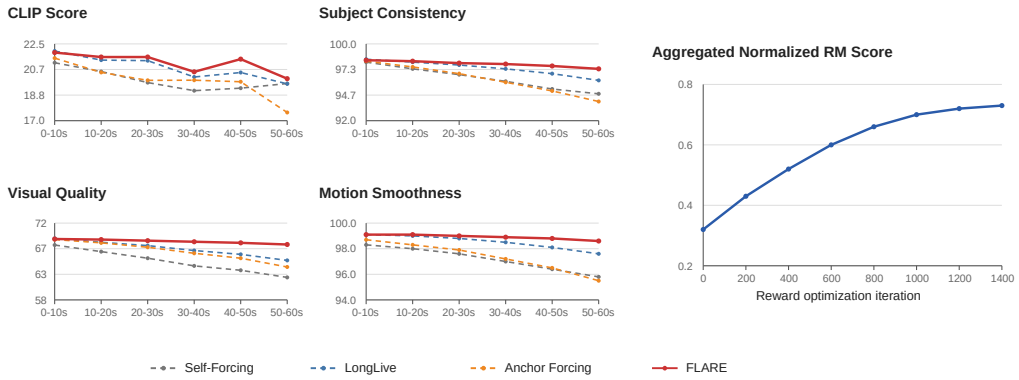

Figure A.1: Chunk-wise long-horizon evaluation and reward score evolution. Left: metrics at different rollout chunks. Compared with autoregressive baselines, FLARE exhibits slower degradation in later chunks. Right: aggregated reward score during generator optimization.

A.4 Interactive Prompt Generation

To construct high-quality multi-segment prompt sequences, we design an automated pipeline powered by a large language model (MiMo-v2.5-pro (Xiaomi, 2026)). For each training sample, the pipeline produces a sequence of six temporally coherent text prompts that describe a continuous shot evolving across segments. Generation proceeds autoregressively: the first segment is synthesized from a randomly sampled story direction and subject type, establishing the main subject, setting, visual style, and camera language. Each subsequent segment is then generated conditioned on both the initial anchor prompt (to preserve scene identity) and the immediately preceding segment (to maintain temporal continuity), with the instruction to introduce exactly one new visible change — either a new action or a new object — for the existing subject (Fig A.2).

We construct preference pairs from 200 prompt groups. Each group contains six ordered prompts, and each 10-second segment is split into two 5-second halves. We pair clips from three long-video autoregressive models with same-prompt clips from two short-video models, yielding $14{,}400$ candidate pairs.

For Consensus-Guided Latent Preference Adaptation, we use multiple pixel-space evaluators as weak annotators and retain only comparisons with head-wise consensus. Tab. A.2 reports the consensus agreement rate for each aspect, computed as the percentage of candidate pairs for which the non-abstaining evaluators predict the same preference direction. The retention rates remain around 70–80% across aspects, indicating that the filtered pairs provide stable supervision for latent reward adaptation.

Prompt of Generating Interactive Prompts

You are a video-prompt generation specialist. Your task:
- Receive an ORIGINAL_PROMPT for the previous part of a continuous shot.
- Write one stand-alone English paragraph of 60-80 words that shows the next moment of the same shot.
- Add exactly one new primary visual change for the existing main subject: either one new action or one new object.
- Convert story-beat wording into concrete visible details. Do not write meta words such as "story beat", "narrative arc", "cue", "trigger", or "closure".
- Keep setting, subject, mood, style, camera scale, and camera movement or angle exactly as in the ORIGINAL_PROMPT.
- Repeat the key subject, clothing, setting, lighting, background, and camera language explicitly, similar to repeated prompt variants in an interactive video dataset.
- Elements may vanish only if naturally obscured by the new action.
- Do not move to a later scene, a new location, a completed final result, or a different camera target.
- Do not use phrases like "still the same", "as before", or "continues" that reveal you read the prior text. Physical stillness such as "lies perfectly still" is allowed when it describes the visible subject.
- You may say the background remains the same, the setting remains unchanged, or the same camera framing is used when this helps visual continuity.
- Use clear mid-level English; avoid rare or literary words.
- Output format MUST be exactly one line, wrapped between <OUTPUT> and </OUTPUT>.
- Do NOT add explanations, greetings, headings, numbering, markdown, or extra lines.
- Anything written outside the two tags will be ignored.

Figure A.2: Prompt template for generating interactive multi-segment prompts.

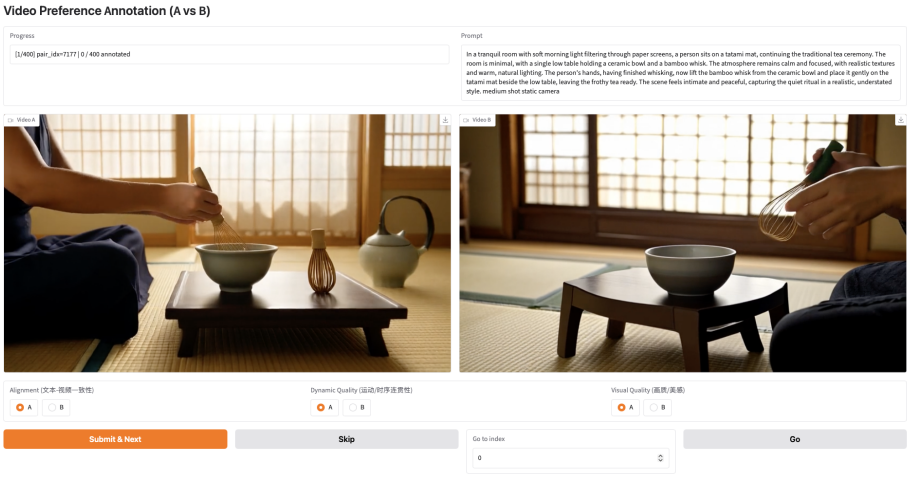

Figure A.3: Internal pairwise preference annotation interface.



0-10s: A focused gardener in their late 30s works alone inside a vast, humid greenhouse, tending to rows of leafy vegetable plants. Sunlight streams through the foggy glass panels, casting a warm, golden glow and creating soft shadows across the wet floor. The air is thick with the smell of damp earth and growing things. With careful, deliberate motions, the gardener uses a small hand tool to gently loosen the soil around the base of a tomato plant, their brow furrowed in concentration. The camera maintains a steady, observational distance, capturing the quiet, meditative rhythm of their task. medium shot handheld micro-shake



10-20s: A focused gardener in their late 30s works alone inside a vast, humid greenhouse, tending to rows of leafy vegetable plants. Sunlight streams through the foggy glass panels, casting a warm, golden glow and creating soft shadows across the wet floor. The air is thick with the smell of damp earth and growing things. With careful, deliberate motions, the gardener uses a small hand tool to gently loosen the soil around the base of a tomato plant, their brow furrowed in concentration, then pauses to gently pluck a small, yellowing leaf from the plant with their fingers. The camera maintains a steady, observational distance, capturing the quiet, meditative rhythm of their task. medium shot handheld micro-shake



20-30s: A focused gardener in their late 30s works alone inside a vast, humid greenhouse, tending to rows of leafy vegetable plants. Sunlight streams through the foggy glass panels, casting a warm, golden glow and creating soft shadows across the wet floor. The air is thick with the smell of damp earth and growing things. After plucking the yellowing leaf, the gardener carefully places the small hand tool into a canvas pouch at their hip, their brow still furrowed in concentration. The camera maintains a steady, observational distance, capturing the quiet, meditative rhythm of their task. medium shot handheld micro-shake



30-40s: A focused gardener in their late 30s works alone inside a vast, humid greenhouse, tending to rows of leafy vegetable plants. Sunlight streams through the foggy glass panels, casting a warm, golden glow and creating soft shadows across the wet floor. The air is thick with the smell of damp earth and growing things. After plucking the yellowing leaf, the gardener carefully places the small hand tool into a canvas pouch at their hip, their brow still furrowed in concentration. The camera maintains a steady, observational distance, capturing the quiet, meditative rhythm of their task. medium shot handheld micro-shake



40-50s: A focused gardener in their late 30s works alone inside a vast, humid greenhouse, tending to rows of leafy vegetable plants. Sunlight streams through the foggy glass panels, casting a warm, golden glow and creating soft shadows across the wet floor. The air is thick with the smell of damp earth and growing things. After plucking the yellowing leaf, the gardener carefully places the small hand tool into a canvas pouch at their hip, their brow still furrowed in concentration. The camera maintains a steady, observational distance, capturing the quiet, meditative rhythm of their task. medium shot handheld micro-shake



50-60s:A focused gardener in their late 30s works alone inside a vast, humid greenhouse, tending to rows of leafy vegetable plants. Sunlight streams through the foggy glass panels, casting a warm, golden glow and creating soft shadows across the wet floor. After misting the tomato plant, the gardener sets the spray bottle down on the wet floor and picks up a pair of small pruning shears from their belt, their brow still furrowed in concentration. The camera maintains a steady, observational distance, capturing the quiet, meditative rhythm of their task. medium shot handheld micro-shake

Figure A.4: Example six-segment interactive prompt sequence and generated rollout.