DELTA FORCING: TRUST REGION STEERING FOR INTERACTIVE AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive streaming video generation requires a causal model to react to evolving event conditions while preserving the visual trajectory it has already established. A fundamental challenge lies in balancing *reactivity* and *stability*: models must respond promptly to new events while maintaining temporal coherence over long horizons. Existing approaches typically distill high-quality bidirectional diffusion models into autoregressive generators and further adapt them via streaming long-tuning. However, they often suffer from identity, layout, and scene drift when event switches. We identify the cause of this failure as *conditional bias*, where the (frozen) teacher can produce denoising guidance that is semantically valid for the new event but incompatible with the student's realized history, thereby causing distribution matching to propagate prompt-consistent yet trajectory-inconsistent modes. We propose Delta Forcing, a transition-calibrated steering objective for a distribution-matching distillation. For each generated chunk, Delta Forcing compares the feature-space delta of the teacher's denoised estimate with the delta of the generator's history-grounded rollout, retaining full teacher-driven distribution matching when the transitions agree and shifting toward trajectory-continuity regularization when they diverge. This adaptive supervision restores the balance between reactivity and stability, enabling coherent long-term generation under dynamic conditions. Extensive experiments demonstrate that Delta Forcing improves long-horizon stability while maintaining responsiveness to evolving events.

1 INTRODUCTION

Interactive streaming video generation is emerging as a fundamental capability for applications such as real-time content creation and world modeling, where visual content must be continuously generated under dynamically evolving event conditions (*e.g.*, prompts, actions, or control signals). Unlike offline generic video generation (Wan et al., 2025; Wu et al., 2025; HaCohen et al., 2024; Seedance et al., 2026; Team et al., 2025; Brooks et al., 2024; Lin et al., 2024; Bao et al., 2024), this setting introduces a fundamental tension between *reactivity* and *stability*: the model must respond instantaneously to new events while simultaneously preserving temporal consistency across events.

To make such systems practical, existing approaches typically adopt a two-stage pipeline. *In stage 1*, high-quality but computationally expensive bidirectional models are distilled into efficient few-step autoregressive diffusion transformer (AR-DiT) generators via forcing-based approaches, such as Diffusion Forcing (Chen et al., 2024), Self-Forcing (Huang et al., 2025a; Yin et al., 2025), and Causal Forcing (Zhu et al., 2026) which commonly trained with distribution matching objectives (Yin et al., 2024b;a). *In stage 2*, methods such as LongLive (Yang et al., 2025) and MemFlow (Ji et al., 2025) further apply *streaming long tuning*, fine-tuning such generators on long sequences with evolving event conditions to enable runtime interactivity.

Despite these advances, a persistent failure mode remains in current pipelines. The generated content, while still semantically aligned with the updated event conditions, gradually drifts in scene layout, object appearance and *etc*. This persists even in memory-aware generators (*e.g.*, MemFlow (Ji et al., 2025)) that are explicitly designed to preserve long-term consistency. Previous studies (Yang et al., 2025; Ji et al., 2025; Huang et al., 2025a; Zhu et al., 2026; Cui et al., 2025) often attribute such degradation to the accumulation of autoregressive generation errors. In contrast, we identify a more fundamental cause: a systematic bias introduced across events by the teacher model during stage 2

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.

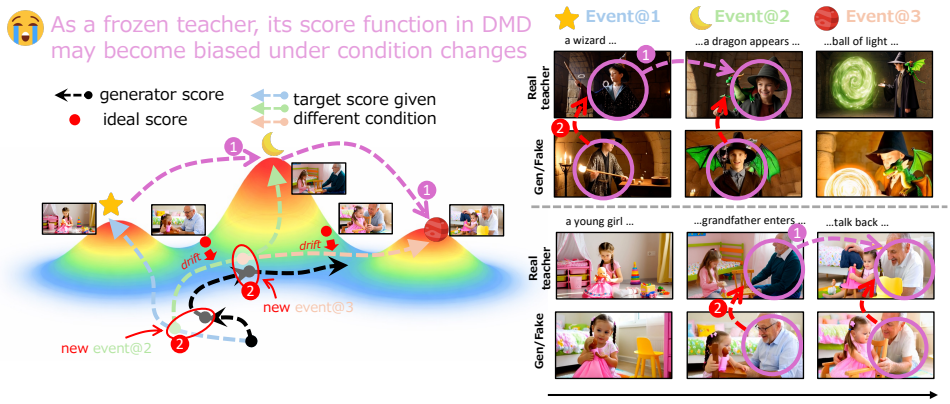

Figure 1: **Left:** Under evolving events, the frozen teacher, biased toward certain patterns, remains condition-aware but trajectory-agnostic, inducing *conditional bias* that deviates from the historical trajectory. **Right:** Decoding both the real teacher model (*i.e.*, Wan2.1-14B-T2V (Wan et al., 2025)) and generator (MemFlow (Ji et al., 2025)) shows that the generator's drift closely follows these teacher-induced shifts.

training. **Existing approaches implicitly assume that the teacher provides reliable supervision under evolving event conditions, and thus enforce alignment with the teacher distribution at each event equally**. However, during stage 2 training, the teacher itself may accumulate errors and exhibit biased behavior, which in turn propagates to the student and induces drift. Empirically, when a new event is introduced (*e.g.*, "...grandfather enters..."→"...talk back..."), the teacher may shift toward a condition-dependent distribution that faithfully reflects the updated semantics but fails to account for the previously generated trajectory (*e.g.*, "dark blue clothes"→"white clothes") as indicated by ❶. Consequently, when such guidance is directly enforced via DMD (Yin et al., 2024b;a) (as indicated by ❷), the model is repeatedly driven toward condition-consistent yet trajectory-inconsistent states. We term this phenomenon ***conditional bias***, where condition-driven shifts in the teacher distribution, coupled with inherent data priors, steer generation toward trajectory-inconsistent modes.

In this work, we propose **Delta Forcing**, a framework that restores the balance between *reactivity* and *stability* under conditional bias. Instead of directly matching the teacher's potentially biased distribution, we shift to *reliability-aware transition modeling* by examining the transition between consecutive latent states. Inspired by trust-region policy optimization (Schulman et al., 2015), where updates are constrained to a region in which the approximation remains reliable, we treat each teacher-induced transition as a candidate update and assess whether it falls within a trajectory-consistent trust region. Concretely, we introduce a delta-based mechanism that gauges the semantic shift of the teacher-induced transition relative to the generator's history-grounded trajectory, and uses this signal to modulate supervision online: transitions within the trust region are reinforced, while those that deviate are suppressed. This adaptive scheme effectively filters out bias-induced guidance, enabling the model to maintain coherent long-term evolution while remaining responsive to new conditions.

In summary, our contributions are threefold. *First*, we identify *conditional bias* in interactive multi-event streaming video generation, a failure mode in which event-driven condition updates steer the teacher toward trajectory-inconsistent modes, corrupting student supervision. *Second*, inspired by trust-region policy optimization, we propose Delta Forcing, a reliability-aware framework that introduces a delta-based mechanism to modulate supervision online, reinforcing trajectory-consistent guidance while suppressing bias-induced deviations. *Third*, extensive experiments demonstrate that Delta Forcing substantially improves temporal stability while preserving event reactivity in long-horizon multi-event scenarios.

2 Related Work

2.1 Stability in Real-time Streaming Video Generation

Real-time streaming video generation requires producing temporally coherent content over long horizons under causal constraints, a setting fundamentally more demanding than offline gen-

eration. A promising paradigm distills a strong pretrained bidirectional DiT into a causal, few-step generator capable of efficient AR synthesis (Huang et al., 2026; Chen et al., 2024; Huang et al., 2025a; Cui et al., 2025; Chen et al., 2026b; Zhu et al., 2026; Liu et al., 2025b; Lu et al., 2025; Yang et al., 2026; Chen et al., 2026a; Liu et al., 2026; Zou et al., 2026). The central challenge in this paradigm is *temporal stability*: errors accumulate over rollouts and manifest as content drift and structural inconsistency, a gap that widens with generation length. A line of work addresses this by narrowing the discrepancy between training and inference. Diffusion Forcing (Chen et al., 2024) trains a causal model with independent per-token noise levels, unifying clean past and noisy future frames to enable stable rollout beyond the training horizon. Self-Forcing (Huang et al., 2025a) closes the train-test gap at its source by conditioning each frame on self-generated rollouts during training. Context Forcing (Chen et al., 2026b) reveals a student–teacher context mismatch in streaming tuning and trains the AR student with a long-context teacher. Causal Forcing (Zhu et al., 2026) identifies that distilling an AR student from a bidirectional teacher violates frame-level ODE injectivity and addresses this by initializing from an AR teacher, substantially recovering generation quality. Reward Forcing (Lu et al., 2025) targets the complementary issue that distilled models suffer diminished motion dynamics, introducing a rewarded distribution matching objective that biases training toward high-reward, high-motion samples. As illustrated in Fig. 2(a), despite their differences, these methods share a common assumption: the student is aligned with teacher guidance under a *fixed* condition. This assumption breaks down in interactive settings where conditions evolve dynamically.

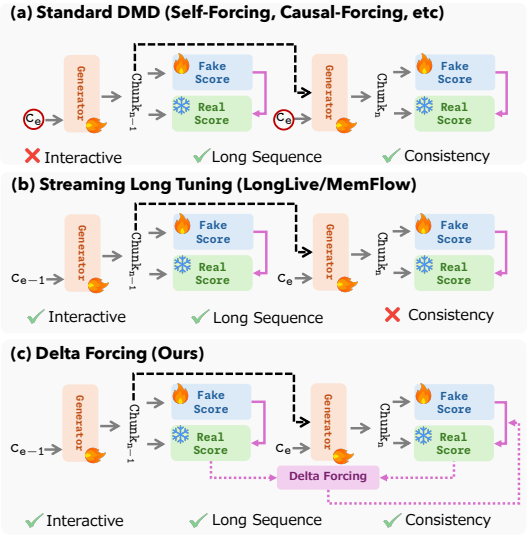

Figure 2: (a) Standard DMD fails to handle condition changes. (b) Streaming Long Tuning improves interactivity but still suffers from biased guidance, and (c) our method enforces transition consistency to mitigate conditional bias and preserve temporal coherence.

2.2 Reactivity in Interactive Streaming Video Generation

A complementary line of work extends AR video generation to interactive settings, where conditions evolve dynamically and the model must adapt to each new event without sacrificing temporal consistency. Early efforts approach interactivity through architectural and scheduling design. SkyReels-V2 (Chen et al.) provides a diffusion-forcing substrate for long rollouts but assumes a fixed, pre-specified condition and does not address mid-rollout event changes. MAGI-1 (Teng et al., 2025) scales AR generation via chunk-wise prediction and exposes chunk-wise prompting as a controllability hook, but each prompt switch requires per-transition tuning of the KV attention range rather than being learned from streaming supervision. A more principled approach targets reactivity through *streaming long tuning*, as illustrated in Fig. 2(b): the generator is fine-tuned on long self-rollouts with evolving conditions, directly aligning training with interactive inference. LongLive (Yang et al., 2025) augments this with a KV-recache mechanism that updates stale context while preserving visual continuity at each event boundary. MemFlow (Ji et al., 2025) replaces memory strategies with prompt-guided dynamic retrieval of relevant historical context, improving narrative coherence across scene switches. Despite these advances, training on evolving conditions does not resolve the underlying supervision problem: at each event boundary, the teacher is steered toward a new condition-dependent distribution that may contradict the established trajectory, providing biased guidance that propagates through subsequent generation.

3 Motivation

Interactive *multi-event* streaming video generation must respond to changing prompts while preserving a continuous, uncut visual trajectory. Existing systems often remain visually plausible and responsive

after a prompt switch, yet abruptly change object identity, appearance, layout, or scene state. The central failure is therefore *cross-event trajectory drift*, not a general collapse of video quality.

Background; We follow the two-stage recipe of LongLive (Yang et al., 2025). Stage 1 distills a bidirectional diffusion teacher into a few-step causal student using distribution matching distillation (DMD) (Yin et al., 2024b;a). Stage 2 applies *streaming long tuning*: the student autoregressively generates long sequences under changing prompts, while DMD is evaluated chunk by chunk. Let $\mathbf{h}_{k-1}$ denote the student's accumulated history, $\mathbf{c}_k$ the current prompt, $\mathbf{x}_k^{\text{new}}$ the new frames, and $\mathbf{o}_{k-1}$ a short overlap copied from the preceding chunk. The effective objective is

$$\mathcal{L}_{\text{stream}} = \mathbb{E}_k \left[\mathrm{KL}\big(p_{\text{fake}}(\mathbf{x}_k^{\text{new}} \mid \mathbf{h}_{k-1}, \mathbf{c}_k) \,\big\|\, p^{\star}_{\text{win}}(\mathbf{x}_k^{\text{new}} \mid \mathbf{o}_{k-1}, \mathbf{c}_k)\big) \right]. \tag{1}$$

The student therefore trains on its own long-horizon rollout, and the frozen teacher already receives the student's realized overlap as local context. In practice, the entire overlap–new-frame window is noised and denoised jointly, with gradients masked to the new frames; Appendix C provides the implementation details. Thus, drift cannot be attributed simply to training without generated history or to a completely context-free teacher.

From observation to an alternative cause. Two properties distinguish the observed failure from ordinary autoregressive degradation. First, the generated video remains visually plausible and follows the updated prompt; what breaks is continuity across the interaction boundary, including identity, appearance, layout, and scene state. Second, drift already appears in the first prompt-transition interval (5–10s), before a long chain of rollout errors can accumulate, and recurs systematically at later prompt switches. Accumulated error may amplify the problem over time, but it does not explain its early onset or its alignment with condition changes. These observations instead suggest that the transition-time supervision itself may introduce the corrupting direction.

Tracing the supervision error. To test this hypothesis, we decode both the frozen teacher's denoised predictions and the student's outputs into pixel space. As shown in Fig. 1(right), the teacher itself changes semantic mode at event boundaries despite receiving the overlap prefix: it produces a visually valid realization of the new prompt but may change an object's identity, appearance, spatial layout, or scene composition. The student subsequently moves in the same semantic direction. Although DMD aligns distributions rather than individual pixels, the synchronized shifts show that the teacher provides a systematic drift direction and that this direction propagates to the student. The latent-trajectory study in Appendix A exhibits the same transition discontinuities.

Root error: conditional bias. The overlap prefix is necessary for local continuation, but it is short and noised; it cannot make a bidirectional teacher trained on short, single-condition clips fully aware of the student's long-horizon trajectory. At a prompt switch, the updated condition can dominate this limited context and pull the teacher toward a condition-consistent mode that conflicts with the established trajectory. We call this residual mismatch *conditional bias*: the teacher remains condition-aware and visually competent, yet its transition is trajectory-inconsistent, and DMD transfers that bias to the student.

From root error to method. Teacher guidance remains useful during stable evolution but becomes unreliable when its transition abruptly departs from the established trajectory. This diagnosis motivates a reliability-aware objective that detects abnormal teacher displacement, suppresses the corresponding DMD update, and falls back on the student's recent trajectory momentum. Inspired by the trust-region principle (Schulman et al., 2015), this leads directly to Delta Forcing in Sec. 4.

4 METHOD

From principle to design. The principle in Sec. 3 leads to two complementary components: a teacher-history gate w_k that suppresses unreliable updates, and a student-history loss $\mathcal{L}_{\text{cont}}$ that preserves recent trajectory momentum. We combine them as

$$\mathcal{L} = w_k \mathcal{L}_{\text{DMD}} + \lambda_{\text{cont}} (1 - w_k)\, s_k\, \mathcal{L}_{\text{cont}}, \tag{2}$$

where λ_{cont} controls the fallback strength and s_k balances the two loss scales. Thus, reliable teacher updates use DMD, while unreliable ones shift weight to student continuity.

Algorithm 1 Delta Forcing

Require: Causal video generator G_θ, prompt set $\mathcal{P}$, video length l_{video}, chunk length l_{chunk}, and overlap length l_{ov}
Require: Delta Forcing: feature extractor Φ, probe level t_d, threshold z_0, sharpness γ_z, clip factor c, and weight λ_{cont}

(a) Streaming Long Tuning

1: Initialize KV cache $C \leftarrow []$ and length $l \leftarrow 0$
2: Initialize sliding-window queue $Q \leftarrow []$ +
3: Sample $(p, p_{\text{next}}) \sim \mathcal{P}$
4: Sample τ_{sw}; $\tau_{\text{sw}} \leftarrow \tau_{\text{sw}} l_{\text{chunk}}$
5: **while** not converged **do**
6: **if** $l \geq l_{\text{video}}$ **then**
7: Reset $C \leftarrow []$, $Q \leftarrow []$, and $l \leftarrow 0$
8: Resample $(p, p_{\text{next}}) \sim \mathcal{P}$
9: Resample τ_{sw}; $\tau_{\text{sw}} \leftarrow \tau_{\text{sw}} l_{\text{chunk}}$
10: **end if**
11: $k \leftarrow l / l_{\text{chunk}} + 1$
12: $p_{\text{active}} \leftarrow \begin{cases} p, & l < \tau_{\text{sw}}, \\ p_{\text{next}}, & \text{otherwise} \end{cases}$
13: **if** $l = \tau_{\text{sw}}$ **then**
14: $C \leftarrow \texttt{recache}(G_\theta, Q, C, p_{\text{active}})$
15: **end if**
16: $\mathbf{x}_{\text{fake}}^{(k)} \leftarrow \texttt{generate_next_chunk}(G_\theta, C, p_{\text{active}})$
17: $\mathbf{W}_k \leftarrow [\text{sg}(Q); \mathbf{x}_{\text{fake}}^{(k)}]$ +
18: $Q \leftarrow \texttt{tail}(\mathbf{W}_k, l_{\text{ov}})$ +
19: $\mathcal{L} \leftarrow \texttt{DMD_Loss}(G_\theta, \mathbf{W}_k, p_{\text{active}})$ −
20: $\mathcal{L}$.backward()
21: Update θ and loss EMAs
22: $l \leftarrow l + l_{\text{chunk}}$
23: **end while**

(b) Delta Forcing: replace Line 19

Persistent state: teacher statistics (μ, σ, n), descriptor caches, and loss EMAs

// Design 1
19a: $\mathbf{f}_k^{\text{F}} \leftarrow \Phi(\mathbf{x}_{\text{fake}}^{(k)})$
19b: $\boldsymbol{\delta}_k^{\text{F}} \leftarrow \mathbf{f}_k^{\text{F}} - \text{sg}(\mathbf{f}_{k-1}^{\text{F}})$
19c: $\mathcal{L}_{\text{cont}} \leftarrow \|\boldsymbol{\delta}_k^{\text{F}} - \text{sg}(\boldsymbol{\delta}_{k-1}^{\text{F}})\|_2^2$
// Design 2
19d: $\hat{\mathbf{x}}_{\text{real}}^{(k)} \leftarrow u_{\text{real}}(\alpha_{t_d} \mathbf{x}_{\text{fake}}^{(k)} + \sigma_{t_d} \boldsymbol{\epsilon}; p_{\text{active}}, t_d)$
19e: $\mathbf{f}_k^{\text{R}} \leftarrow \Phi(\hat{\mathbf{x}}_{\text{real}}^{(k)})$
19f: $d_k^{\text{R}} \leftarrow \|\mathbf{f}_k^{\text{R}} - \mathbf{f}_{k-1}^{\text{R}}\|_2$; $z_k \leftarrow (d_k^{\text{R}} - \mu)/\sigma$
19g: $w_k \leftarrow \sigma(-\gamma_z(z_k - z_0))$
19h: $(\mu, \sigma, n) \leftarrow \texttt{update}\big(\min(d_k^{\text{R}}, \mu + c\sigma)\big)$
// Scale balancing
19i: $\mathcal{L}_{\text{DMD}} \leftarrow \texttt{DMD_Loss}(G_\theta, \mathbf{W}_k, p_{\text{active}})$
19j: $s_k \leftarrow \overline{\mathcal{L}}_{\text{DMD}} / \overline{\mathcal{L}}_{\text{cont}}$
// Delta Forcing objective
19k: $\mathcal{L} \leftarrow w_k \mathcal{L}_{\text{DMD}} + \lambda_{\text{cont}}(1 - w_k) s_k \mathcal{L}_{\text{cont}}$

↪ **Continue with Lines 20–23 in (a).**

Observation 1. *The student trajectory provides an endogenous momentum signal when the current teacher update is unreliable.* Following the chunk-wise autoregressive paradigm of LongLive (Yang et al., 2025), the generator produces $\mathbf{x}_{\text{fake}}^{(k)} \sim G_\theta(C_k, \mathbf{c}_{e(k)})$, where C_k is the accumulated KV cache encoding the previously generated trajectory and $e(k)$ denotes the event index at chunk k. Because C_k is constructed from the generator's own previous outputs, the temporal evolution of the fake samples records the direction along which the student trajectory has recently evolved. This history-grounded signal provides a local momentum prior without requiring an additional external target.

Design 1: Student momentum preservation. For compactness, let $\mathbf{f}_k^{\text{F}} = \Phi(\mathbf{x}_{\text{fake}}^{(k)})$ denote the semantic descriptor of the student output. We define

$$\boldsymbol{\delta}_k^{\text{F}} = \mathbf{f}_k^{\text{F}} - \text{sg}(\mathbf{f}_{k-1}^{\text{F}}), \quad \mathcal{L}_{\text{cont}} = \|\boldsymbol{\delta}_k^{\text{F}} - \text{sg}(\boldsymbol{\delta}_{k-1}^{\text{F}})\|_2^2. \tag{3}$$

Here, Φ is a frozen DINO feature extractor (Oquab et al., 2023; Siméoni et al., 2025), with latent decoding and preprocessing omitted for clarity, and $\text{sg}(\cdot)$ denotes stop-gradient on cached history. Unlike directly minimizing $\|\mathbf{f}_k^{\text{F}} - \mathbf{f}_{k-1}^{\text{F}}\|_2^2$, which suppresses first-order motion, Eq. equation 3 penalizes changes between consecutive displacements. It therefore preserves a non-zero but locally consistent student momentum.

Observation 2. *An unreliable teacher update can be detected as an outlier relative to the teacher's own temporal variation.* The frozen teacher may naturally exhibit different levels of semantic variation across videos and motion patterns. Consequently, a fixed threshold on the absolute teacher displacement cannot reliably distinguish conditional bias from legitimate dynamics. We instead maintain persistent statistics of the teacher's semantic displacements and detect whether the current displacement is a standardized outlier with respect to this baseline.

Crucially, teacher reliability is determined from teacher self-consistency rather than teacher–student agreement. Around an event transition, the student may legitimately follow a trajectory different from the teacher's current denoised estimate. Using their disagreement as the reliability criterion would therefore confound teacher bias with ordinary model mismatch.

Design 2: History-calibrated teacher reliability. Fresh noise levels and classifier-free guidance make the DMD teacher output unsuitable for temporal comparison. We instead run a guidance-free, fixed-noise probe at every chunk (no gradients; $\sim$20% step-time overhead):

$$\hat{\mathbf{x}}_{\text{real}}^{(k)} = u_{\text{real}}\Big(\alpha_{t_d}\mathbf{x}_{\text{fake}}^{(k)} + \sigma_{t_d}\boldsymbol{\epsilon};\, \mathbf{c}_{e(k)}, t_d\Big)\,, \qquad \mathbf{f}_k^{\text{R}} = \Phi\Big(\hat{\mathbf{x}}_{\text{real}}^{(k)}\Big). \tag{4}$$

Using the normalized descriptor, we standardize consecutive teacher displacement and convert it into reliability:

$$d_k^{\text{R}} = \left\|\mathbf{f}_k^{\text{R}} - \mathbf{f}_{k-1}^{\text{R}}\right\|_2, \qquad z_k = \frac{d_k^{\text{R}} - \mu_k}{\sigma_k}, \qquad w_k = \sigma(-\gamma_z(z_k - z_0))\,. \tag{5}$$

The persistent statistics (μ_k, σ_k) use Welford warm-up followed by an exponential moving average, with displacements clipped at $\mu_k + c\,\sigma_k$ before each update. Thus, a temporal outlier closes the gate and activates student momentum regularization, while typical teacher evolution retains DMD supervision.

Scale balancing. Because $\mathcal{L}_{\text{DMD}}$ and $\mathcal{L}_{\text{cont}}$ have different raw scales, we balance them using their exponential moving averages:

$$s_k = \frac{\overline{\mathcal{L}}_{\text{DMD}}}{\overline{\mathcal{L}}_{\text{cont}}}. \tag{6}$$

This makes λ_{cont} the scale-independent relative strength of the fallback regularizer.

Delta Forcing. Equation equation 2 and Alg. 1 combine the two branches: historically typical teacher transitions retain standard DMD supervision ($w_k \approx 1$), whereas transition-time displacement spikes downweight the teacher and activate the student's momentum regularizer. Persistent legitimate dynamics are absorbed into the running statistics, allowing DMD to recover. Further analysis and ablations appear in Appendix B, Sec. 6, and Appendix D.

5 Experiments

5.1 Implementation Details

Implementation. We build Delta Forcing on top of LongLive (Yang et al., 2025). The generator is WAN-2.1-1.3B-T2V (Wan et al., 2025) and the teacher is WAN-2.1-14B-T2V (Wan et al., 2025), which natively produces 5-second clips at 16 FPS and 832×480 resolution. Training proceeds in two stages as LongLive (Yang et al., 2025). In Stage 1, we replace the original Self Forcing (Huang et al., 2025a) initialization with Causal Forcing (Zhu et al., 2026) for the context window and frame sink, and train for 700 steps with learning rates of 2×10^{-6} for the generator and 4×10^{-7} for $\texttt{S}_{\texttt{fake}}$. In Stage 2, we apply Delta Forcing together with *Streaming Long Tuning* for an additional 3,000 steps, with learning rates of 1×10^{-5} and 2×10^{-6}, respectively.

Baselines. We compare Delta Forcing against two families of methods at a comparable scale: open-source causal video generators, including SkyReels-V2 (Chen et al.) and MAGI-1 (Teng et al., 2025), and distilled causal video generators including LongLive (Yang et al., 2025), MemFlow (Ji et al., 2025), and Reward Forcing (Lu et al., 2025). For a fair comparison, all distilled causal baselines are initialized following the Causal Forcing (Zhu et al., 2026), and Reward Forcing is further trained under *Streaming Long Tuning* to enhance interactive generation capability.

Evaluation. We evaluate Delta Forcing along four complementary axes. ❶ We adopt VBench (Huang et al., 2024; Zheng et al., 2025; Huang et al., 2025b) to assess overall performance. Specifically, we adopt six dimensions: Subject Consistency, Background Consistency, Aesthetic Quality, Motion Smoothness, Dynamic Degree, and Imaging Quality. ❷ We measure instruction-following ability using Long-CLIP (Zhang et al., 2024). ❸ We further evaluate human-preference alignment with VideoAlign (Liu et al., 2025a), which scores visual quality (VQ), motion quality (MQ), and text alignment (TA). ❹ Finally, we conduct a user study to obtain direct human preference evaluation. All experiments are conducted on the MemFlow benchmark (Ji et al., 2025), which contains 100 sequences, each consisting of six sequential 10-second events and forming a 60-second video.

Table 1: Comparisons with Long-CLIP Score (Radford et al., 2021) and VideoAlign (Liu et al., 2025a) metrics across different video generation baselines.

Model	Long-CLIP Score ↑							VideoAlign ↑			
	0-10s	10-20s	20-30s	30-40s	40-50s	50-60s	**Avg.**	VQ	MQ	TA	**Total**
Diffusion Forcing Causal											
SkyReels-V2 (Chen et al.)	27.08	25.29	24.42	24.25	23.88	23.79	24.79	2.24	3.45	-0.20	5.49
MAGI-1 (Teng et al., 2025)	27.15	25.25	23.73	23.67	23.40	23.61	24.47	1.95	1.53	-0.83	2.66
Distilled Causal											
LongLive (Yang et al., 2025)	27.69	26.39	25.68	**25.53**	25.35	25.66	26.05	2.67	3.96	0.50	7.13
MemFlow (Ji et al., 2025)	27.27	26.47	25.44	25.36	25.27	25.13	25.82	**2.83**	4.10	0.54	7.46
Reward Forcing (Lu et al., 2025)	27.58	25.04	24.35	23.95	23.80	23.75	24.75	2.72	4.18	0.04	6.94
Delta Forcing (Ours)	**28.11**	**26.55**	**26.01**	25.52	**25.51**	**25.71**	**26.24**	2.79	**4.19**	**0.57**	**7.55**

5.2 Quantitative Results

❶ Results on VBench.

As shown in Table 2, Delta Forcing attains the best subject and background consistency within the distilled-causal group while matching competing methods on motion smoothness, aesthetic quality, and imaging quality. The large gains on transition-sensitive consistency without trading off perceptual or motion quality indicates that the improvements come from more stable cross-event trajectory evolution under repeated condition updates rather than from a shift along any single perceptual axis. Methods optimized for instantaneous responsiveness remain highly reactive but accumulate transition mismatch over long horizons. Our trust-region supervision directly targets this trade-off by suppressing trajectory-incompatible teacher updates across events.

Table 2: Comparisons with related representative open-source autoregressive video generation models of similar parameter sizes on VBench.

Model	#Params	Evaluation Scores ↑					
		Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Dynamic Degree
Diffusion Forcing Causal							
SkyReels-V2	1.3B	91.64	92.88	98.46	56.97	64.33	88.37
MAGI-1	4.5B	92.63	91.20	98.43	59.18	65.12	82.04
Distilled Causal							
LongLive	1.3B	94.97	93.37	98.37	58.95	72.59	90.93
MemFlow	1.3B	93.19	92.03	97.17	55.91	69.57	91.71
Reward Forcing	1.3B	95.55	93.40	98.51	56.32	**74.99**	**93.16**
Delta Forcing	1.3B	**96.63**	**94.82**	**98.71**	**58.95**	73.11	92.32

❷ Delta Forcing maintains instruction-following capability. Since commonly adopted prompt-enhancement strategies (Wan et al., 2025; Wu et al., 2025) typically produce prompts that are more detailed and substantially longer than standard CLIP-style captions (Radford et al., 2021), we measure prompt–video alignment for each interaction chunk using Long-CLIP (Zhang et al., 2024). This metric directly evaluates whether the model can follow newly arriving prompts as the conditioning signal evolves over time. As shown in Table 1, Delta Forcing remains consistently top-ranked or near top-ranked from early to late windows, indicating that transition-aware supervision preserves responsiveness while improving long-horizon stability.

❸ Delta Forcing achieves stronger human preference. We further evaluate Delta Forcing using VideoAlign (Liu et al., 2025a), a reward model trained from human feedback, to better capture perceptual quality and human preference in interactive generation. To make the evaluation boundary-sensitive, we concatenate each generated interaction chunk with the final 1-second segment of the preceding interaction and score the resulting stitched clip. This protocol explicitly emphasizes event continuity, where conditional bias is most likely to induce drift. As shown in Table 1, Delta Forcing achieves the best overall VideoAlign score, with the highest motion quality and text alignment, while remaining competitive in visual quality.

❹ User Study. We conduct a side-by-side user study with 20 participants to assess quality in multi-event streaming generation. In each trial, participants view four videos generated under the same sequence of event-driven prompts and rank them along three criteria: aesthetic quality, dynamic quality, and multi-event naturalness. Because this ranking is cognitively demanding, each participant completes only four trials, randomly sampled from the full set. Video order within each trial is randomized. All methods are evaluated on identical prompt schedules, enabling direct comparison of both responsiveness to condition changes and temporal consistency across event boundaries. The



Figure 3: Qualitative results. Each 10s segment corresponds to one event and the full event prompts are listed in Appendix E.

user-study interface is shown in Appendix F. As summarized in Table 3, Delta Forcing attains the best (lowest) average rank across all three criteria, indicating a clear human preference. The margin is largest on dynamic quality and multi-event naturalness, consistent with our method's design goal of preserving coherent motion and producing smoother transitions under evolving conditions.

5.3 Qualitative Results

Figure 3 presents qualitative comparisons for multi-event generation. The regions highlighted by purple circles reveal the failure patterns in existing methods: the generated content drifts across events, manifesting as changes in object identity, scene layout, and overall structure. This clearly indicates that the trajectory is not smoothly evolved, but instead repeatedly redirected toward different modes after each event. Although MemFlow (Ji et al., 2025) incorporates memory mechanisms that can partially recover from drift (*e.g.*, in Fig. 3, Example 1, the appearance returns closer to the initial state from 40–50s to 50–60s), it still fails to maintain consistent trajectories due to biased supervision. This suggests that memory alone is insufficient when the teacher guidance itself is not history-aware. In contrast, Delta Forcing, despite being built upon a memoryless generator following LongLive (Yang

Table 3: **User Study.** Evaluation results on human preference via average rank.

Model	Aesthetic ↓	Dynamic ↓	Naturalness* ↓	Average ↓
LongLive	2.78	2.44	2.63	2.62
MemFlow	2.68	3.08	2.89	2.88
Reward Forcing	2.41	2.55	2.74	2.57
Delta Forcing (Ours)	**2.13**	**1.93**	**1.74**	**1.93**

*Naturalness is evaluated from a multi-event perspective, measuring the smoothness and coherence of transitions across events.



Knob	Setting	Subject Consist.	Backgr. Consist.	Motion Smooth.	Dynamic Degree
default	z_0=1.0, γ_z=1.5, λ=0.5	**96.63**	**94.82**	**98.71**	**92.32**
threshold z_0	0.5 (early gating)	96.74	94.90	98.82	87.95
	2.0 (rare gating)	95.12	93.28	98.41	92.87
sharpness γ_z	0.75 (soft gate)	96.21	94.12	98.69	91.96
	3.0 (near-binary)	96.35	94.24	98.12	91.42
fallback λ_{cont}	0.25	95.88	93.81	98.52	92.43
	1.0	96.63	94.58	98.80	89.61

Figure 4: **Left:** Qualitative component ablations. **Right:** Quantitative sensitivity analysis of the gate hyperparameters.

et al., 2025), maintains consistent object identity, structure, and motion across events. The generated trajectory evolves smoothly without abrupt deviations, demonstrating that our method effectively suppresses biased teacher guidance and preserves trajectory-consistent dynamics. We further provide latent-trajectory visualizations in Appendix A.4 to empirically validate Delta Forcing's effectiveness.

6 Ablation Studies

Without continual loss (Design 1). Removing Design 1 leaves the adaptive trust region on the DMD objective. Local appearance and motion remain stable (Fig. 4, row 3), showing that the gate still rejects some unreliable teacher guidance. Nevertheless, the global layout gradually drifts. The gate decides whether to trust the teacher, but cannot provide a student-history target when that trust weakens. Thus, the continual loss serves as a trajectory anchor rather than a visual-quality regularizer.

Without adaptive trust regions (Design 2). Removing Design 2 applies DMD supervision uniformly and produces persistent upward camera motion (row 4). This shows that the continual fallback alone cannot cancel a consistently biased, mode-seeking gradient; it only stabilizes local evolution. Similar camera-panning degeneration under dynamics-oriented guidance has been reported by Reward Forcing (Lu et al., 2025). Together, the ablations establish complementary roles: Design 1 preserves accumulated trajectory momentum, while Design 2 filters unreliable teacher updates.

Sensitivity analysis. Fig. 4 (right) varies one gate hyperparameter at a time under the same training budget and 100-case evaluation. The threshold endpoints fail oppositely: early gating (z_0=0.5) over-stabilizes motion (dynamic degree 92.32→87.95), whereas rare gating (z_0=2.0) re-admits drift (subject/background consistency 95.12/93.28). Sharpness is less sensitive: both endpoints remain close to the default on consistency, although a near-binary gate (γ_z=3.0) reduces motion smoothness from 98.71 to 98.12 through abrupt supervision switches. The fallback weight exhibits the expected trade-off: weak anchoring (λ_{cont}=0.25) lowers consistency, while strong anchoring (λ_{cont}=1.0) dampens dynamics from 92.32 to 89.61. These trends match the intended roles of z_0, γ_z, and λ_{cont}: they control intervention frequency, transition sharpness, and fallback strength, respectively. No variant improves both stability and dynamics over the default, supporting its balanced operating point. Structural ablations of persistent statistics and scale balancing appear in Appendix D.

7 Conclusion

We identify *conditional bias* as a structural failure mode in distillation-based streaming video generation: the teacher is condition-aware but trajectory-agnostic, so distribution-matching supervision repeatedly steers the student toward condition-consistent but trajectory-inconsistent modes, accumulating drift across event boundaries. We propose Delta Forcing, a reliability-aware framework inspired by trust-region policy optimization that compares each teacher-induced transition with the generator's history-grounded evolution, reinforcing supervision within a trajectory-consistent trust region and suppressing guidance that deviates from it. Extensive experiments show that Delta Forcing substantially improves temporal stability while preserving event reactivity in long-horizon real-time interactive multi-event video generation.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript language editing, organization, and LaTeX formatting. All AI-assisted material was reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of this work, including its text, claims, and artifacts.

ETHICS STATEMENT

This work studies interactive video generation, a technology that can be misused to create deceptive or harmful media and may inherit biases from pretrained models and their data. Our method does not address these risks; real-world deployment should include provenance, access control, and content-safety safeguards. Our evaluation includes a 20-participant study in which participants ranked anonymized synthetic model outputs; participants were not asked to provide personal media, and only aggregate rankings are reported. The authors have read and adhered to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 4 and Algorithm 1 specify the method and training procedure, while Section 5 reports the model configurations, optimization schedule, baselines, benchmark, and evaluation metrics. The appendix further documents the overlapping-window implementation (Appendix C), structural ablations (Appendix D), benchmark construction and prompts (Appendix E), and user-study protocol and interface (Appendix F). Together, these materials provide the details needed to reproduce the training and evaluation setup.

REFERENCES

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion, December 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: infinite-length film generative model (2025). *URL https://arxiv. org/abs/2504.13074*.

Jintao Chen, Chengyu Bai, Xinda Xue, Mu Xu, et al. Grounded forcing: Bridging time-independent semantics and proximal dynamics in autoregressive video synthesis. *arXiv preprint arXiv:2604.06939*, 2026a.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026b.

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. *arXiv preprint arXiv:2510.02283*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. Live: Long-horizon interactive video world modeling. *arXiv preprint arXiv:2602.03747*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion, November 2025a.

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.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. doi: 10.1109/TPAMI.2025.3633890.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. MemFlow: Flowing adaptive memory for consistent and efficient long video narratives, December 2025.

Jiaxing Li, Kai Zou, Cindy Zhou, Kaichen Huang, Junyao Gao, Zile Wang, Yang Liu, Bin Liu, Bo An, and Yangguang Li. Distillalign: Coordinating mode covering and mode seeking in autoregressive video distillation. *arXiv preprint arXiv:2607.26811*, 2026.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

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*, 2025a.

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-HsuanYang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation, 2026. URL https://arxiv.org/abs/2603.09488.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time, September 2025b.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward Forcing: Efficient Streaming Video Generation with Rewarded Distribution Matching Distillation, December 2025.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

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.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation, October 2025.

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*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step Diffusion with Distribution Matching Distillation, October 2024b.

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, September 2025.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

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, February 2026.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Nenghai Yu. Hiar: Efficient autoregressive long video generation via hierarchical denoising. *arXiv preprint arXiv:2603.08703*, 2026.

A MOTIVATION STUDY VIA LATENT TRAJECTORY VISUALIZATION

To supplement our motivation analysis, we provide a latent-space diagnostic that reveals how existing interactive streaming video generation methods behave under evolving event conditions. Concretely, we extract the denoised latent feature of every generated frame prior to VAE decoding and visualize the temporal trajectory of these latent states in a two-dimensional PCA space. This analysis offers direct evidence of whether the generation process evolves smoothly across prompt switches or instead exhibits drift, collapse, or mode-seeking behavior.

A.1 WHAT SHOULD A GOOD TRAJECTORY LOOK LIKE?

Before presenting the visualization protocol and results, we first establish the criteria against which trajectories should be evaluated. From the perspective of interactive multi-event video generation, a well-behaved latent trajectory should satisfy two complementary properties:

- **Within-interaction aggregation.** Frames belonging to the same prompt interval should form a compact local cluster, reflecting stable identity, scene layout, and motion style. Such aggregation indicates that the model maintains a coherent visual state while generating temporally continuous frames under a fixed condition.
- **Smooth cross-interaction displacement.** When the prompt changes, the trajectory should move from the previous local region toward a new one in a gradual and structured manner. The displacement should be sufficient to reflect the semantic shift, yet not so abrupt as to break temporal continuity.

In other words, a desirable trajectory should resemble a sequence of compact local groups connected by smooth bridges. This structure reflects the central trade-off of our problem: within-interaction stability and between-interaction responsiveness must coexist. Excessively small displacement implies insufficient adaptation, whereas abrupt jumps indicate loss of temporal consistency. With these criteria in mind, we now describe the visualization protocol and examine whether existing methods satisfy them.

A.2 VISUALIZATION IMPLEMENTATIONS

Latent extraction and projection. For each model, we generate a video under the same multi-event interaction setting in which the prompt changes at fixed intervals. Suppose the video contains T frames divided into K interaction segments, with each segment corresponding to one sub-prompt. Let $\mathbf{z}_t \in \mathbb{R}^d$ denote the denoised latent representation of frame t before VAE decoding. We collect the full latent sequence:

$$\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T\}. \tag{7}$$

We fit PCA on $\mathcal{Z}$ and project each frame-level latent into a two-dimensional space:

$$\tilde{\mathbf{z}}_t = \mathbf{W}^\top(\mathbf{z}_t - \boldsymbol{\mu}), \qquad \tilde{\mathbf{z}}_t \in \mathbb{R}^2, \tag{8}$$

where $\boldsymbol{\mu}$ is the empirical mean of the sequence and $\mathbf{W} \in \mathbb{R}^{d\times 2}$ contains the top two principal directions. The projected points $\{\tilde{\mathbf{z}}_t\}_{t=1}^T$ are connected in temporal order to form a trajectory, with different colors denoting different interaction segments. We additionally annotate the start and end of selected segments to facilitate reading of transition patterns. To ensure fair comparison, all methods share the same prompt schedule, switching timestamps, and latent extraction stage.

Choice of projection method. We adopt PCA rather than t-SNE (van der Maaten & Hinton, 2008) or UMAP (McInnes et al., 2018) because our objective is not cluster discovery but *temporal trajectory analysis*. This distinction has three concrete implications. *First*, PCA preserves the *global geometry* of the latent space. As a linear projection along the directions of maximal variance, it maintains relative distances and directional relationships. t-SNE and UMAP prioritize local neighborhood fidelity and can substantially distort global arrangement, which is acceptable for cluster visualization but problematic when assessing whether a sequence follows a coherent trajectory. *Second*, PCA faithfully reflects *transition characteristics*. Because it preserves large-scale directional variation, the projected trajectory allows us to inspect transition magnitude, smoothness, and cross-event

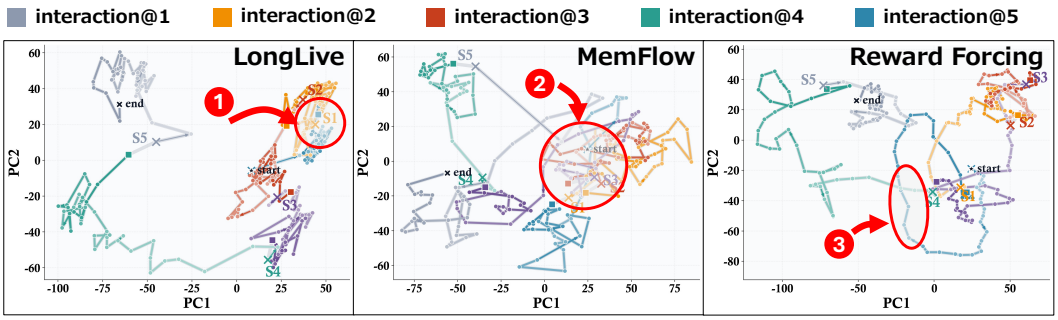

Figure 5: **Latent trajectory visualization via PCA under multi-event prompt switching.** We project frame-wise denoised latent features (before VAE decoding) into a two-dimensional PCA space and connect them in temporal order. Different colors denote different interaction segments. Left exhibits short and narrow transitions across prompt switches, indicating insufficient semantic displacement despite changed conditions. Middle: shows irregular and scattered patterns with no clear temporal structure, indicating severe inter-event drift. Right evolves with persistent motion regardless of interaction boundaries, suggesting mode-seeking dynamics in which within-event and cross-event transitions become indistinguishable. Comparing against the desirable properties defined in Section A.1, none of these methods achieves both within-interaction aggregation and smooth cross-interaction displacement.

displacement without the artificial stretching or compression that nonlinear methods may introduce. A smooth trajectory that appears fragmented, or an abrupt jump that appears mild, would directly undermine our diagnostic purpose. *Third*, PCA is *deterministic and reproducible*. t-SNE and UMAP are sensitive to perplexity, neighborhood size, initialization, and random seed. Since our conclusions depend on comparing trajectory patterns across methods under identical conditions, we require a projection whose behavior is stable across runs.

A.3 Empirical Observations

Figure 5 presents the PCA trajectories of three representative methods under consecutive prompt switching. We evaluate each against the desirable properties established in Section A.1.

❶ **Under-reactive transitions.** The trajectory exhibits short, concentrated transitions across the first several prompt segments. Despite clear shifts in the semantic condition, the latent path displaces only modestly, remaining anchored to preceding states rather than moving toward the new event condition. With respect to our criteria, this method partially satisfies within-interaction aggregation but fails on cross-interaction displacement: the trajectory appears temporally smooth, yet the displacement is too weak to reflect the actual semantic change. This pattern indicates an under-reactive transition mode in which continuity is preserved at the cost of insufficient adaptation.

❷ **Unstructured drift.** The trajectory exhibits a markedly different failure. Rather than forming a structured progression, latent states scatter across the projected space without directional regularity. Consecutive prompt transitions correspond to abrupt jumps, and the overall path lacks coherent temporal organization. This method violates both desirable properties: within-interaction frames do not form compact clusters, and cross-interaction transitions are neither smooth nor controlled. The behavior is consistent with severe inter-event drift, where the model is repeatedly pulled toward different condition-consistent modes without maintaining a stable connection to the prior trajectory.

❸ **Mode-seeking persistence.** The trajectory continues moving along a similar pattern regardless of whether frames belong to the current interaction or a new one, making within-event and cross-event transitions largely indistinguishable. Against our criteria, this method fails primarily on within-interaction aggregation: instead of forming compact local clusters for each prompt interval, the trajectory is dominated by a persistent motion tendency—particularly camera motion—that overrides event-conditioned structure. The model follows an internally favored motion pattern rather than reorganizing around each new event, constituting a form of mode-seeking dynamics.

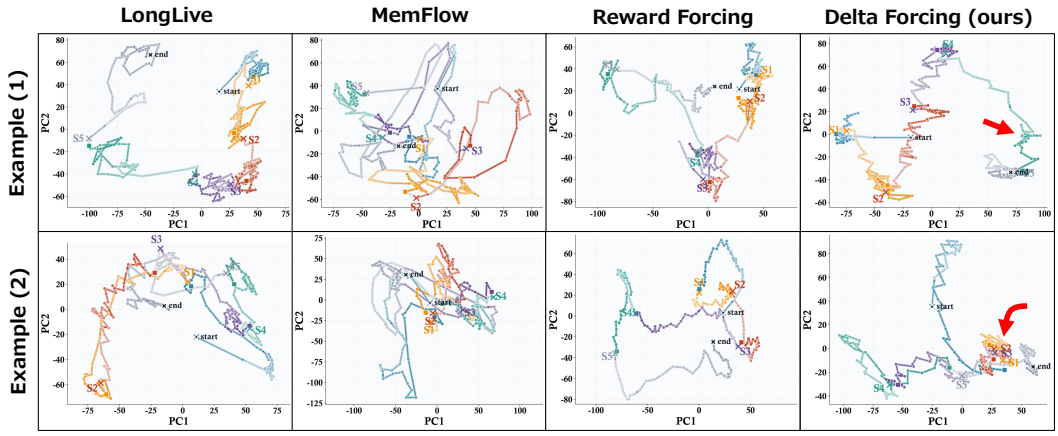

Figure 6: **Extended latent trajectory comparison.** Each row shows one example under the same multi-event prompt schedule, comparing three baselines (columns 1–3) against Delta Forcing (column 4). Red arrows highlight segments where Delta Forcing exhibits compact within-interaction clusters connected by smooth cross-interaction transitions, consistent with the desirable properties established in Section A.1.

A.4 Further Latent Trajectory Comparisons

To verify that the trajectory patterns identified in Section A.3 are consistent beyond a single instance, we extend the PCA-based latent trajectory analysis to additional examples and include Delta Forcing for comparison. Figure 6 presents two representative cases (rows), each comparing three baselines against Delta Forcing (rightmost column) under the same multi-event prompt schedule.

Across both examples, the baseline methods reproduce the failure modes previously identified. These patterns are consistent with the under-reactive, unstructured-drift, and mode-seeking behaviors discussed in Section A.3, confirming that these failure modes are systematic rather than instance-specific.

In contrast, Delta Forcing (rightmost column, red arrows) produces trajectories that more closely satisfy the desirable properties defined in Section A.1. Within each interaction segment, the latent states remain relatively aggregated, forming compact local clusters. Across prompt switches, the trajectory transitions smoothly to a new region with sufficient displacement to reflect the semantic change. This combination of intra-state stability and responsive inter-state transitions provides latent-space evidence that Delta Forcing better balances within-interaction coherence and between-interaction adaptability.

B Distinguishing Trajectory Drift from Legitimate Dynamics

A potential concern is whether a large teacher displacement necessarily indicates trajectory drift, since legitimate object or camera motion may also induce substantial feature changes. Our method does not classify drift based on the absolute displacement magnitude. Instead, it detects whether the current displacement is a standardized outlier relative to the persistent statistics of the teacher's own displacement stream.

Different temporal signatures. Legitimate dynamics are generally temporally coherent. For example, a fast-moving object or sustained camera motion may produce consistently large teacher displacements over several consecutive chunks. Although d_k^{R} can be large in absolute value, it remains within the typical range of the displacement statistics:

$$z_k = \frac{d_k^{\mathrm{R}} - \mu_k}{\sigma_k} \approx 0. \tag{9}$$

Consequently, the gate stays open and the corresponding DMD supervision is preserved.

Trajectory drift exhibits a different temporal pattern in our setting. We focus on one-shot interactive generation, where the video is expected to evolve continuously without intentional shot cuts. In our experiments, teacher-induced drift is predominantly observed around prompt-transition boundaries. Moreover, it usually manifests as a rapid semantic or structural jump, rather than a slowly accumulating degradation. At a transition chunk k_e, this produces

$$z_{k_e} \gg z_0, \tag{10}$$

where (μ, σ) are determined by the pre-transition displacement history. Such a localized standardized spike is unlikely to arise from the continuous evolution expected in a one-shot video and is therefore treated as an unreliable teacher update. Because displacements are clipped at $\mu + c\,\sigma$ before entering the statistics, the spike does not contaminate its own reference baseline.

Behavior around prompt transitions. When an abrupt teacher jump is detected, the reliability weight w_{k_e} decreases and suppresses the corresponding DMD update. Meanwhile, the complementary weight $1 - w_{k_e}$ increases the contribution of the student delta-consistency loss. This prevents the student from immediately following the teacher's discontinuous transition and instead encourages it to preserve its recent trajectory momentum. Therefore, the method does not suppress the new prompt itself. It suppresses the impulsive manner in which the teacher attempts to realize the prompt at the transition boundary.

Adaptation to legitimate new dynamics. A new prompt may legitimately introduce a different motion pattern. If this change persists across subsequent chunks, its displacements become temporally consistent rather than remaining isolated outliers. As the new (clipped) displacements are absorbed into the persistent statistics by the exponential moving average, (μ_k, σ_k) adapt to the new dynamic regime and z_k decreases. Consequently, w_k increases and DMD supervision is restored within a few tens of observations. The gate therefore mainly affects the short transition interval:

$$\underbrace{z_{k_e} \gg z_0}_{\text{transient boundary jump}} \quad \longrightarrow \quad w_{k_e} \approx 0, \tag{11}$$

while, after adaptation,

$$\underbrace{z_{k_e+r} \approx 0}_{\text{sustained new dynamics}} \quad \longrightarrow \quad w_{k_e+r} \approx 1. \tag{12}$$

Role of semantic features. We compute the displacement in the DINO feature space rather than directly in pixel space. This further reduces sensitivity to local appearance changes, pixel-level noise, and ordinary object motion, while retaining sensitivity to larger semantic and structural changes. The gate consequently responds primarily to unexpected trajectory-level changes instead of raw motion magnitude.

Scope of the detector. The proposed gate should be understood as a detector for abrupt, transition-induced drift rather than a universal classifier for all forms of generation error. A legitimate motion that begins instantaneously at exactly the prompt boundary may also be temporarily downweighted. However, because persistent motion is rapidly absorbed into the persistent displacement statistics, this suppression is short-lived. Conversely, a very gradual drift that never produces a local displacement outlier may not be directly detected. Our design targets the dominant failure mode observed in one-shot prompt-switching generation: a rapid trajectory jump caused by the frozen teacher immediately responding to a new condition without respecting the previously generated history.

C Would a History-Aware Teacher Be the More Natural Fix?

Our diagnosis of conditional bias invites a natural counter-proposal: if the failure stems from a teacher that does not condition on the student's realized history, the direct fix is a *history-aware teacher*—for instance, replacing the T2V teacher with an image- or video-conditioned model such as Wan2.1-I2V-14B during Stage-2 supervision. We agree with the diagnosis behind this proposal but not with the instantiation. This section makes three points: (i) our streaming long-tuning recipe *already* supervises the student through a history-conditioned teacher, in a soft and symmetric form that is effectively a video-conditioned continuation (V2V) teacher, (ii) making the conditioning explicit and asymmetric, as with an off-the-shelf I2V teacher, is structurally counterproductive, and (iii) the transition failures

that motivate Delta Forcing persist under any windowed history conditioning, because they originate in the text-conditioning pathway rather than in the absence of history.

Streaming long tuning already trains through a continuation teacher. Following LongLive's streaming long tuning (Yang et al., 2025), our Stage-2 training is chunk-wise autoregressive with *overlapping windows*. Each training window is a fixed 21-latent-frame chunk assembled as

$$\mathbf{W}_k = \big[\underbrace{\mathrm{sg}(\mathbf{x}_{k-1}^{\mathrm{tail}})}_{\text{overlap prefix}};\ \underbrace{\mathbf{x}_k^{\mathrm{new}}}_{\text{generated}}\big], \tag{13}$$

where the prefix is the detached tail of the previous chunk (typically 3 latent frames, the number of new frames is re sampled every step) and the first frame of the assembled window is re-encoded as an image latent, matching the inference-time `recache` convention. The student generates the new frames on its persistent KV cache, so it is a genuine continuation student: its training-time generation distribution coincides with its streaming inference distribution. The DMD loss is then computed on the *full* window: all 21 frames are noised at a shared timestep, the frozen teacher denoises the whole window, and gradients are masked to the newly generated frames. Consequently, the teacher's estimate for the new frames is conditioned on the (noised) realized history prefix—the supervision is effectively $\nabla \log p(\mathbf{x}^{\mathrm{new}} \mid \mathbf{x}_{k-1}^{\mathrm{tail}}, c)$, *i.e.*, a video-conditioned continuation teacher, obtained without any architectural modification.

Why this soft, symmetric form is the right one. Three properties distinguish it from explicit I2V conditioning. *Symmetry:* the fake score is evaluated on exactly the same window, so the prefix-conditioning contribution appears in both terms of the DMD score difference and cancels, the update remains a proper divergence gradient over the new frames. *Softness:* the prefix enters the teacher at the same noise level as the new frames, so the anchoring strength decays gracefully at high noise instead of acting as a hard clean-still anchor. *In-distribution:* the teacher performs exactly the windowed denoising it was trained for—no bespoke conditioning pathway, no inputs it has never seen.

Why an explicit I2V teacher is structurally counterproductive. Instantiating the reviewer's proposal literally—a frozen I2V teacher conditioned on the student's boundary frame as the real score, with the student and critic unchanged—breaks all three properties at once, and each break is independently harmful. *(a) Asymmetric conditioning.* Only the real score receives the anchor image $\mathbf{I}_{k-1}$: it estimates the conditional $\nabla \log p(\mathbf{x} \mid c, \mathbf{I}_{k-1})$ while the critic estimates the student's marginal $\nabla \log q(\mathbf{x} \mid c)$. Their difference retains a residual "anchor attraction" term that the critic can never learn to cancel, and the T2V student has no mechanism to represent the conditioning explicitly, the mode-seeking update resolves every disagreement by pulling the student toward copying the anchor. *(b) A self-referential target.* The anchor is produced by the student itself, so the distillation target $p(\cdot \mid c, f(q))$ moves with the student and admits degenerate fixed points: any output the teacher would faithfully continue into itself—frozen content included—is self-consistent, allowing the loss to stay small while the student leaves the data manifold. *(c) A low-entropy conditional under reverse KL.* An I2V model concentrates its mass on continuations that stay close to the anchor, distilling toward this low-entropy conditional under mode-seeking reverse KL systematically rewards near-copy behavior. This is the extreme case of the coverage contraction analyzed by DistillAlign (Li et al., 2026): the target is swapped mid-training from the T2V marginal (which the Stage-1 student was initialized to cover) to a per-chunk conditional family whose dominant modes are near-static—the coverage collapse that reverse KL induces gradually in the standard setting becomes the primary training signal. Compounding all three, the teacher is queried at conditioning inputs that are out-of-distribution (generated frames, further degraded by the VAE decode–encode round trip), turning it from a restoring force into an error amplifier: artifacts entering through the conditioning image are endorsed rather than corrected, so perturbations compound across chunks instead of being damped. A prototype of this baseline in our codebase exhibited exactly this signature—progressive loss of motion followed by content collapse, while training losses remained healthy—consistent with the analysis.

What windowed history conditioning cannot fix. The overlap prefix spans a single 21-frame window, so the teacher's history awareness is local by construction. At a prompt switch, the window's denoising target is dominated by the *text* pathway: the frozen teacher, now conditioned on c_{e+1}, proposes realizations of the new event that are prompt-consistent but incompatible with the prefix—which is precisely the conditional bias characterized in Sec. 3. History conditioning of the windowed

form is therefore necessary but not sufficient: it stabilizes within-event continuation yet leaves transition-time supervision biased. Delta Forcing targets exactly this residual failure, acting on the *loss* (displacement-gated trust in the teacher) rather than on the conditioning pathway.

D Structural Ablation Studies

The sensitivity analysis over the gate hyperparameters (z_0, γ_z, λ_{cont}) is presented in the main paper (Fig. 4, Sec. 6). Here we additionally ablate the two structural choices of the method: the automatic scale balancing s_k (Eq. equation 6, replaced by a fixed s_k=1) and the persistent-statistics gate itself (replaced by a per-sequence sliding-window baseline). Each variant changes exactly one component from the default configuration and is trained with an identical schedule and budget on the same data, all variants are evaluated on the full 100-case interactive benchmark. Both structural components matter, and they fail in the directions the method analysis predicts: without automatic scale balancing, the raw continuity loss (roughly an order of magnitude larger than the DMD term) over-weights the momentum branch on gated steps and dampens motion (dynamic degree 92.32→89.27), while the per-sequence sliding-window gate spends most of each sequence in warm-up, misses transition spikes, and consequently re-admits part of the drift the persistent gate removes (subject/background consistency 96.63/94.82→95.63/93.60).

Table 4: **Structural ablations.** Removing the automatic scale balancing (fixed s_k=1) and replacing the persistent z-score gate with the per-sequence sliding-window baseline.

Variant	Subject Consist.	Backgr. Consist.	Motion Smooth.	Dynamic Degree
Delta Forcing (full)	**96.63**	**94.82**	**98.71**	**92.32**
w/o scale balancing	96.42	94.35	98.74	89.27
sliding-window gate	95.63	93.60	98.49	91.75

E Experiment Details

Benchmark Description. This benchmark evaluates a video generation model's ability to produce long-form, narratively coherent video by decomposing a one-minute scene into six consecutive 10-second segments. Unlike single-shot text-to-video benchmarks that evaluate isolated clips, this benchmark stresses the model's capacity for cross-event consistency: maintaining stable character identity, environmental continuity, and a progressive storyline as the timeline unfolds.

Formulation. Each test case is a sequence of prompts $\{P_1, P_2, \ldots, P_N\}$ where each prompt P_i corresponds to a fixed-length window $[t_i, t_i + \Delta t)$ with $\Delta t = 10$ seconds. The full benchmark sample spans $N \cdot \Delta t = 60$ seconds with $N = 6$ segments. Adjacent prompts P_i and P_{i+1} describe **consecutive events** in the same continuous scene — there are no cuts, scene changes, or temporal jumps between segments. Persistent entities (characters, setting, key objects) are explicitly re-described in every prompt to anchor visual identity, while the **action and state evolve incrementally**.

Example@1 — Young Wizard & Dragon

`00:00 – 00:10`: A young wizard with a pointed hat and a sparkling robe is practicing magic in an ancient, stone-walled chamber. He waves a wooden wand, causing small objects on a table to float in the air. The chamber is lit by flickering torchlight, casting long shadows.

`00:10 – 00:20`: A young wizard with a pointed hat in an ancient, stone-walled chamber continues to levitate objects with his wand. A small, scaly dragon with shimmering green wings flies into the chamber through a high window and lands on his shoulder. The wizard looks surprised but pleased.

`00:20 – 00:30`: A young wizard with a pointed hat and a small green dragon on his shoulder practice magic together in the ancient, stone-walled chamber. He points his wand at a

dusty book, and the dragon lets out a tiny puff of smoke, causing the book to flip open. The torchlight flickers on the walls.

00:30 – 00:40: A young wizard with a pointed hat and a small green dragon on his shoulder in the ancient, stone-walled chamber laugh as the dragon sneezes, sending a shower of harmless sparks onto the floating objects. The wizard pats the dragon's head affectionately.

00:40 – 00:50: A young wizard with a pointed hat and a small green dragon on his shoulder in the ancient, stone-walled chamber work together to create a larger spell. The wizard chants an incantation while the dragon flaps its wings, and a ball of swirling light forms between them.

00:50 – 01:00: A young wizard with a pointed hat and a small green dragon on his shoulder in the ancient, stone-walled chamber watch as the ball of light floats up to the ceiling and bursts into a shower of magical stars. They look at each other and share a triumphant smile.

Example@2 — Park Ranger & Fawn

00:00 – 00:10: A park ranger in a green uniform is walking along a dirt trail in a vast, sun-drenched national park. The landscape is filled with towering redwood trees and ferns. The air is clear and crisp.

00:10 – 00:20: The park ranger walking in the vast national park discovers a small, orphaned fawn huddled at the base of a giant redwood tree. The fawn looks weak and scared. The forest is quiet.

00:20 – 00:30: The park ranger and the small, orphaned fawn are in the vast national park. The ranger kneels down slowly, speaking in a soft, calm voice to avoid frightening the animal further. The redwood trees tower over them.

00:30 – 00:40: The park ranger in the vast national park gently scoops up the small, orphaned fawn into their arms. The fawn is light and fragile. The sun filters through the massive redwood canopy.

00:40 – 00:50: The park ranger in the vast national park carries the small, orphaned fawn carefully, cradling it against their uniform as they begin the long walk back to the ranger station. The forest path is dappled with sunlight.

00:50 – 01:00: The park ranger walks with purpose through the vast national park, a protector of the forest and its creatures, carrying the small, orphaned fawn to safety. The giant redwood trees stand as silent witnesses.

F User Study

Figure 7 shows the web-based interface used for our human evaluation. For each trial, participants are shown a single text prompt at the top, followed by four videos generated by the compared models (anonymized as Model A–D, with order randomized per trial). The videos are displayed side-by-side with synchronized playback controlled by shared *Play* and *Restart* buttons, allowing direct visual comparison. Below the videos, an evaluation guide briefly defines the three criteria: **Aesthetic Quality** (visual appeal, composition, lighting, and clarity), **Dynamic Quality** (meaningfulness and prompt-consistency of motion), and **Multi-Event Naturalness** (smoothness of transitions across narrative stages while preserving identity consistency). Participants are instructed to evaluate each criterion independently rather than producing a single overall judgment. For each criterion, they rank the four models from best to worst by dragging labeled chips into an ordered slot row, and submit rankings per trial.

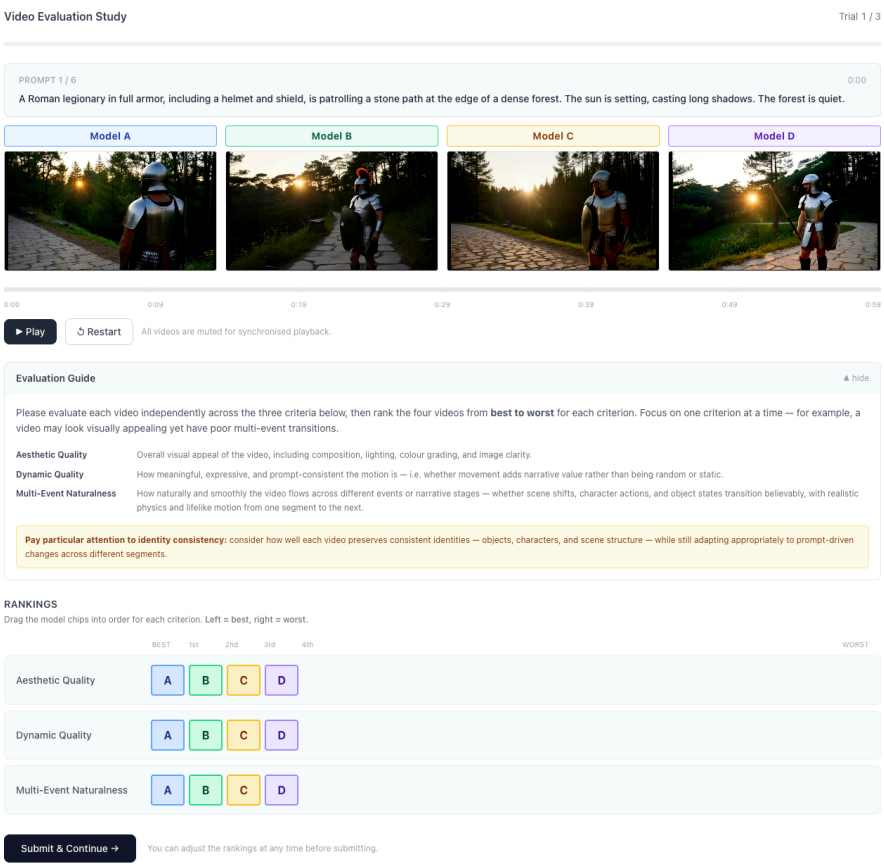

Figure 7: User study interface.