CONTEXT PERSISTENCE CALIBRATION FOR INTERACTIVE LONG VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive long video generation entails adapting to successive prompts while preserving coherent content over extended horizons. Causal autoregressive models are well suited to this setting, as sequential generation and KV cache reuse naturally support streaming synthesis. Existing methods primarily preserve historical context through visual KV caches, while paying comparatively limited attention to the persistence of textual conditioning. Yet textual conditioning provides direct semantic guidance by specifying the desired visual content. In interactive settings, the initial prompt often establishes global attributes such as subject appearance, scene, and visual style. As generation proceeds under successive prompts, relying primarily on cached visual states to preserve these global attributes becomes increasingly unreliable, degrading long-range visual consistency and semantic coherence. To address this limitation, we propose Context Persistence Calibration (CPC), a training-free framework that maintains global semantics while accommodating evolving instructions. Specifically, we jointly contextualize the initial and current prompts and retain the leading singular components of the contextualized initial-prompt representation. We term this procedure *Semantic Anchor*, which constructs a rank-constrained textual reference at each subsequent interaction stage. To further mitigate *sink-collapse* and the resulting abrupt scene resets, we introduce *Adaptive Sink Regulation*, which adaptively regulates excessive contributions from the frame sink. CPC operates entirely at inference time without additional training, learnable parameters, or cache expansion. Experiments on two interactive long video evaluation sets, InterVidProM and InterStoryBench, demonstrate improved long-range visual consistency and semantic coherence, reduced *sink-collapse*, and continued responsiveness to evolving prompts.

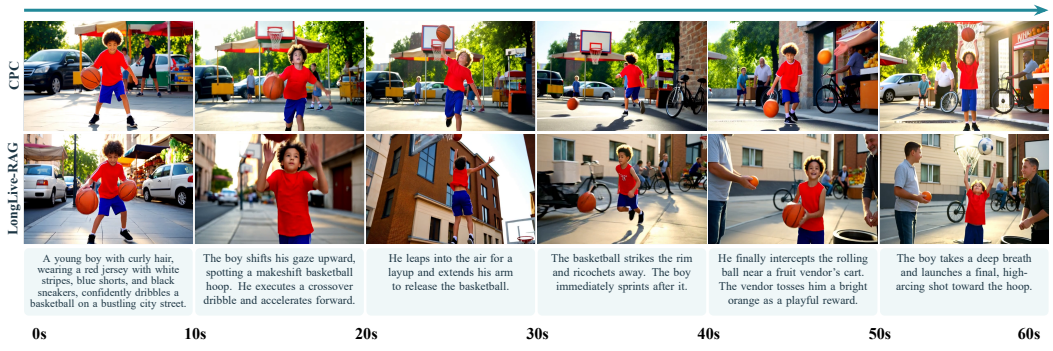

Figure 1: Interactive long video generation with CPC. CPC generates a continuous 60-second video from six sequential prompts, with each prompt controlling a 10-second segment, while maintaining long-range visual consistency across prompt transitions.

1 INTRODUCTION

Video diffusion models have achieved substantial progress in visual quality, temporal coherence, and text-guided controllability (Jin et al., 2025; Kong et al., 2024; Polyak et al., 2024; Wan et al., 2025).

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.

Recent advances in causal autoregressive video diffusion have further extended video synthesis to long-horizon and interactive settings, where streaming prompts progressively modify actions, states, viewpoints, or scenes (Brooks et al., 2024; Yang et al., 2025a). Maintaining visual consistency and semantic coherence while adapting to successive prompts becomes increasingly difficult over long horizons, as autoregressive errors accumulate and the conditioning signal changes across interaction stages.

Long video generation has been explored through diverse generation paradigms. Diffusion-based video generation models and Diffusion-Forcing models achieve strong visual quality and temporal coherence (Wei et al., 2025; Yang et al., 2025b; Chen et al., 2026b). Meanwhile, causal autoregressive models support efficient sequential generation through causal attention and KV cache reuse, making them well suited to long-horizon and streaming video generation (Ge et al., 2022; Yu et al., 2024; Yin et al., 2025). To sustain such long-horizon autoregressive generation under bounded computation and memory, existing methods retain early visual states as persistent frame sinks and maintain recent states through rolling KV caches, providing long-range references and local temporal context, respectively (Li et al., 2026a; Huang et al., 2025a). Existing work also explores positional extrapolation (Li et al., 2025; Yesiltepe et al., 2026), historical-state compression (Yi et al., 2025; Yu et al., 2025), and retrieval of relevant content from earlier video segments (Hu et al., 2026) to support longer generation horizons.

Despite these advances, existing methods focus primarily on maintaining visual context through the KV cache, while paying comparatively limited attention to the persistence of textual conditioning. Yet textual conditioning remains important for specifying the desired semantic content of the generated video. In interactive generation, the initial prompt typically establishes global attributes such as subject appearance, scene configuration, and visual style, while subsequent prompts specify evolving actions, states, viewpoints, or scene changes (see Fig. 1). Following a prompt transition, however, the initial prompt is replaced by the subsequent instruction, leaving its global specifications to be preserved primarily through cached visual states. Such preservation may become unreliable over extended autoregressive generation, as accumulated errors can gradually weaken the representation of these specifications in the cached visual context. Consistent with this analysis, Fig. 2 shows the video–text similarity of the generated frames to both the initial prompt and its sequence-invariant description [1]. The similarity to both textual references progressively declines across interaction stages. Notably, the decline with respect to the sequence-invariant description indicates reduced adherence to persistent visual attributes that should remain unchanged despite evolving actions or viewpoints. Although a straightforward remedy is to directly concatenate the initial prompt with the current instruction, this can introduce semantic interference between outdated stage-specific content and the evolving instruction, leading to repeated early-stage actions or weakened adherence to the current prompt (see Figs. 9–10 in Appendix D).

In this paper, we propose Context Persistence Calibration (CPC), a training-free framework that maintains explicit access to the initial textual context across prompt transitions while accommodating evolving instructions. Specifically, we jointly contextualize the initial and current prompts and retain the leading singular components of the contextualized initial-prompt representation. This procedure, termed *Semantic Anchor*, constructs a rank-constrained textual reference that persists across subsequent interaction stages, complementing recursively generated visual history with explicit textual guidance while preserving responsiveness to evolving instructions. Complementing this textual calibration, we also address *sink-collapse*, a long-horizon autoregressive failure mode where generated content reverts toward the sink frame, resulting in abrupt scene resets or cyclic motion patterns (Cui et al., 2026). Our analysis in Fig. 4 reveals localized surges in frame-sink attention that coincide with such resets. Motivated by this observation, we further introduce *Adaptive Sink Regulation*, which selectively regulates excessive frame-sink contributions. CPC operates entirely at inference time without additional training or learnable parameters and does not increase the visual KV-cache capacity.

To evaluate CPC under successive prompt transitions, we construct two interactive long video evaluation sets, InterVidProM and InterStoryBench, derived from VidProM (Wang & Yang, 2024) and StoryBench (Bugliarello et al., 2023), respectively. Extensive experiments demonstrate that CPC

[1] Extracted from the initial prompt by comparing it with subsequent prompts and retaining only the visual attributes that remain valid throughout the sequence.

improves long-range visual consistency and semantic coherence, reduces scene resets, and remains responsive to evolving prompts. Our main contributions are summarized as follows:

- We identify insufficient persistence of the initial textual context across prompt transitions as an important limitation in interactive long video generation.
- We propose *Context Persistence Calibration* (CPC), a training-free framework that constructs a persistent, rank-constrained textual reference through joint contextualization and singular-value truncation, while adaptively regulating excessive frame-sink contributions.
- We construct InterVidProM and InterStoryBench for evaluating interactive long video generation, and extensive experiments demonstrate the effectiveness of CPC across different generation horizons and interaction settings.

2 Related Work

Long-Horizon and Interactive Video Generation. Long-horizon video generation has advanced through diffusion forcing, autoregressive diffusion, causal distillation, and self-generated rollouts (Chen et al., 2025; Sand.ai et al., 2025; Yin et al., 2025; Huang et al., 2025a; Cui et al., 2025; Liu et al., 2025). Interactive generation supports prompt updates via chunk-wise conditioning, KV recache, and planner–executor control (Sand.ai et al., 2025; Yang et al., 2025a; Wang et al., 2026).

Long-Range Context Management. Strategies for managing long-range visual context include frame compression (Zhang & Agrawala, 2025), slow-fast memory (Chen et al., 2026c), and head-specific cache policies (Chen et al., 2026a). Other approaches explore camera-aware or reconstructed memory (Hong et al., 2025; Sun et al., 2025) and latent retrieval (Hu et al., 2026). Positional extrapolation (Zhao et al., 2025; Yesiltepe et al., 2026) further extends generation horizons, while FLEX combines RoPE modulation, antiphase noise sampling, and inference-time sinks (Li et al., 2026b). We further incorporate textual context into long-range context management by constructing a rank-constrained reference to the initial prompt for interactive generation.

Attention Sinks. StreamingLLM retains initial-token states for stable streaming inference (Xiao et al., 2024), while video models similarly preserve frame sinks for long-range continuity (Yang et al., 2025a; Li et al., 2026a). Deep Forcing adjusts temporal RoPE and compresses history (Yi et al., 2025), Reward Forcing updates sink states to reduce initial-frame copying (Lu et al., 2025), and LoL mitigates sink-collapse with multi-head RoPE jitter (Cui et al., 2026). We adaptively regulate excessive frame-sink attention using the oldest block in the rolling KV cache as a reference, without altering temporal positional encoding.

3 Method

Overview. In this section, we first introduce the generation setup in Sec. 3.1 and analyze context persistence in Sec. 3.2. We then present Context Persistence Calibration (CPC) in Sec. 3.3, which maintains a rank-constrained textual reference across prompt transitions and adaptively regulates the contribution of the frame sink to improve long-horizon generation stability. The overall framework is illustrated in Fig. 5.

3.1 Preliminary

A causal autoregressive video diffusion model generates N temporal video chunks, with the joint distribution factorized as

$$p(x_{1:N} \mid \mathbf{E}^{\mathrm{act}}_{1:N}) = \prod_{n=1}^{N} p(x_n \mid x_{<n}, \mathbf{E}^{\mathrm{act}}_n), \tag{1}$$

where x_n denotes the n-th video chunk, $x_{<n}$ denotes the generated history and $\mathbf{E}^{\mathrm{act}}_n$ is the active textual condition, which remains unchanged within an interaction stage and is updated at prompt transitions. Each chunk is generated through a causal diffusion process with KV cache reuse.

Following LongLive (Yang et al., 2025a), we maintain a bounded KV cache consisting of one persistent frame sink and a rolling KV cache composed of three temporal blocks, denoted by C_1, C_2, and C_3 from oldest to newest. The frame sink retains the initial visual states, while the rolling KV cache is continuously updated with recently generated states. At each prompt transition, KV recache recomputes the rolling KV cache from the generated video history under the new prompt, while leaving the persistent frame sink unchanged.

3.2 Context Persistence Analysis in Interactive Long Video Generation

Degradation of Initial Textual Context. Interactive long video generation is driven by successive prompts. The initial prompt typically establishes global specifications such as subject appearance, scene configuration, and visual style, whereas subsequent prompts introduce evolving actions, states, viewpoints, or scene changes. At each interaction stage, the active prompt is independently encoded as $\mathbf{E}_i = T_\phi(P_i)$, where T_ϕ denotes the frozen T5 text encoder (Raffel et al., 2020). After the first prompt transition these specifications are no longer explicitly from P_1 to P_2, available in the textual condition. The model must then rely primarily on visual states retained in the frame sink and rolling KV cache to preserve them. As generation proceeds, accumulated errors can gradually weaken their representation in the cached context.

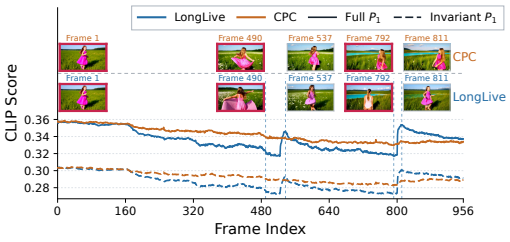

Figure 2: Per-frame CLIP similarity to the complete initial prompt (Full P_1, solid) and its sequence-invariant description (Invariant P_1, dashed). Colors distinguish LongLive and CPC. Burgundy-highlighted regions indicate semantic drift, while CPC better preserves alignment with persistent visual attributes.

To quantify the persistence of the initial textual context in the generated video, we compute the per-frame CLIP score (Radford et al., 2021) with respect to Full P_1 and Invariant P_1, as shown in Fig. 2. We construct Invariant P_1 using Qwen3.6-Plus (Qwen Team, 2026) by using subsequent prompts as references to identify and extract from P_1 the persistent attributes that remain valid throughout the sequence. These persistent attributes mainly include subject appearance and visual style, while transient actions, viewpoints, and other stage-specific content are excluded. Importantly, Invariant P_1 is used only for offline analysis in Fig. 2 and is not part of CPC inference. During generation, CPC accesses only the initial prompt P_1 and the current prompt P_i at each interaction stage.

Both LongLive and CPC exhibit a gradual decline in CLIP Score as generation proceeds, which can partly reflect the evolving visual content across interaction stages. However, the decline is more pronounced for LongLive with respect to both Full P_1 and Invariant P_1, suggesting weaker alignment with the initial textual context, particularly with visual attributes that remain valid across later interaction stages. This trend is also reflected in the generated frames: by Frame 792, for example, the subject's bright pink dress has drifted into a light-colored swimsuit. In comparison, CPC maintains higher alignment with the persistent visual attributes while accommodating intended changes in pose, viewpoint, and scene content. In addition to this gradual trend, the abrupt increases around Frames 537 and 811 correspond to *sink-collapse*.

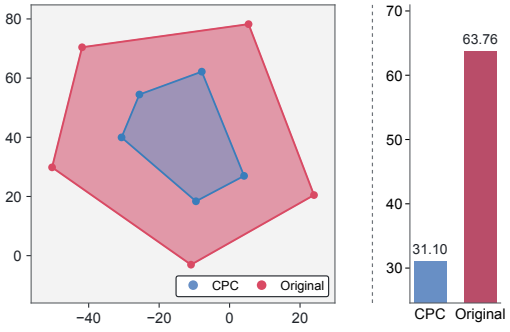

Figure 3: t-SNE visualization of P_2–P_6 textual representations from a six-stage sequence. Original denotes independent prompt encoding and CPC denotes the textual calibration. Left: 2D t-SNE projection. Right: mean pairwise distance in the projected space, shown as a descriptive statistic of the visualization.

We further visualize the textual representations produced by the original prompt encoding and CPC. We visualize only stages P_2–P_6, since CPC is applied after the initial interaction stage. In the original setting, each prompt P_i is encoded independently, whereas CPC jointly processes the

initial prompt P_1 and each current prompt P_i. As shown in Fig. 3, the CPC representations appear more compact in the 2D t-SNE projection, with the mean pairwise distance decreasing from 63.76 to 31.10. Since t-SNE is used only for visualization and does not preserve high-dimensional distances exactly, we treat this result as a qualitative indication of a more compact representation organization rather than as quantitative evidence of semantic preservation. This observation is consistent with the output-side analysis in Fig. 2, where CPC maintains stronger alignment with persistent visual attributes across interaction stages.

Attention Allocation to the Frame Sink. Long-horizon autoregressive generation is also susceptible to *sink-collapse*, in which generated content abruptly reverts toward the sink frames, resulting in scene resets or cyclic motion patterns. To examine this behavior, we track the block-level attention mass allocated to the frame sink and rolling KV cache blocks throughout generation. As shown in Fig. 4, the frame-sink attention mass generally decreases as generation progresses and remains relatively low, except for localized surges around chunks 44 and 67. These surges coincide with scene resets from the evolving content in Frames 504 and 792 toward the initial hiking scene in Frames 536 and 811, respectively. This temporal correspondence suggests that excessive attention allocation to the frame sink is closely associated with *sink-collapse*. In contrast, CPC suppresses these localized surges and preserves the evolving visual trajectory.

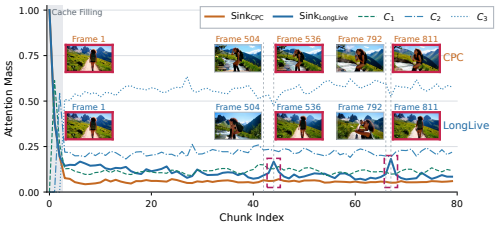

Figure 4: Block-level attention mass allocated to the frame sink and rolling KV cache blocks across chunks, averaged over query tokens and attention heads. The C_1, C_2, and C_3 curves show the original attention masses of the three rolling KV cache blocks for reference. Frame-sink attention surges in LongLive coincide with scene resets, while CPC suppresses surges.

3.3 Context Persistence Calibration

Building on the above analysis, we propose *Context Persistence Calibration* (CPC), which preserves the original frame-sink and rolling KV cache structure while calibrating textual and visual context. As shown in Fig. 5, *Semantic Anchor* constructs a rank-constrained textual reference from the initial and current prompts, while *Adaptive Sink Regulation* selectively regulates excessive frame-sink attention. Our full algorithm is presented in Algorithm 1.

Semantic Anchor. At the first interaction stage, we use the initial prompt without modification. For each subsequent stage $i > 1$, we concatenate the initial prompt P_1 with the current prompt P_i as $\widetilde{P}_i = P_1 \oplus P_i$, where $\oplus$ denotes textual concatenation.

We then use the T5 text encoder T_ϕ to obtain the jointly contextualized representation $\widetilde{\mathbf{E}}_i$. Joint encoding allows the initial and current prompts to interact within a shared contextual space, such that the representation of the initial prompt is contextualized by the current instruction rather than reused as a fixed historical embedding.

For $i > 1$, let $\mathcal{I}_i^{\text{init}}$ and $\mathcal{I}_i^{\text{cur}}$ denote the token regions corresponding to the initial and current prompts. We extract

$$\widetilde{\mathbf{E}}_i^{\text{init}} = \widetilde{\mathbf{E}}_i\left[\mathcal{I}_i^{\text{init}}, :\right], \qquad \widetilde{\mathbf{E}}_i^{\text{cur}} = \widetilde{\mathbf{E}}_i\left[\mathcal{I}_i^{\text{cur}}, :\right]. \tag{2}$$

Directly retaining the complete contextualized representation of P_1 would repeatedly introduce the full initial instruction into subsequent interaction stages, potentially overemphasizing outdated stage-specific information and interfering with the current instruction. We therefore constrain its representational capacity by applying singular value decomposition (SVD) to $\widetilde{\mathbf{E}}_i^{\text{init}} \in \mathbb{R}^{L\times d}$:

$$\widetilde{\mathbf{E}}_i^{\text{init}} = \mathbf{U}_i \operatorname{diag}(\boldsymbol{\sigma}_i)\mathbf{V}_i^\top, \qquad \sigma_{i,1} \geq \cdots \geq \sigma_{i,r}, \tag{3}$$

where $r = \min(L, d)$. Given a retention ratio ρ, we retain $k = \max(1, \lfloor \rho r \rfloor)$ singular components and reconstruct the textual reference as

$$\widehat{\mathbf{E}}_i^{\text{init}} = \mathbf{U}_i \operatorname{diag}\left(\sigma_{i,1}, \ldots, \sigma_{i,k}, 0, \ldots, 0\right) \mathbf{V}_i^\top. \tag{4}$$

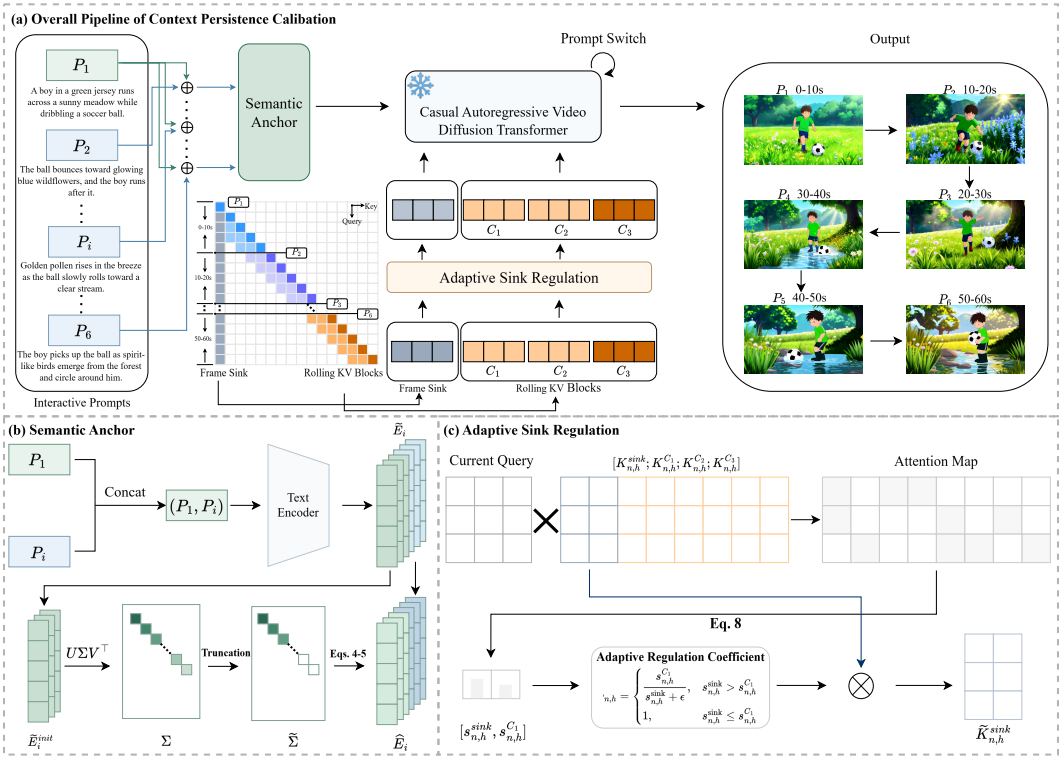

Figure 5: (a) The overall pipeline of Context Persistence Calibration (CPC). Sequential prompts drive causal autoregressive generation, while CPC maintains access to the initial textual context and regulates the frame-sink contribution across prompt transitions. (b) During *Semantic Anchor*, we jointly encode the initial and current prompts and retain the leading singular components to construct a rank-constrained textual reference. (c) In *Adaptive Sink Regulation*, we use C_1 as a dynamic reference to regulate excessive frame-sink attention mass on a per-head basis.

The resulting rank-constrained reference maintains explicit access to the initial textual context while limiting its representational capacity.

Finally, we form the calibrated textual condition by setting $\widehat{\mathbf{E}}_i = \widetilde{\mathbf{E}}_i$ and independently weighting the initial textual reference and the current instruction:

$$\widehat{\mathbf{E}}_i[\mathcal{I}_i^{\text{init}}, :] = \lambda_{\text{init}} \widehat{\mathbf{E}}_i^{\text{init}}, \qquad \widehat{\mathbf{E}}_i[\mathcal{I}_i^{\text{cur}}, :] = \lambda_{\text{cur}} \widetilde{\mathbf{E}}_i^{\text{cur}}. \tag{5}$$

Here, λ_{init} and λ_{cur} weight the initial reference and current instruction, respectively. For each stage $i > 1$, the reference is reconstructed according to the current prompt and $\widehat{\mathbf{E}}_i$ replaces the independently encoded $\mathbf{E}_i$ as the textual condition.

Adaptive Sink Regulation. At autoregressive step n, for each attention head h, we consider the full cached context consisting of the frame sink and three rolling KV cache blocks: $\mathbf{K}_{n,h}^{\text{cache}} = [\mathbf{K}_{n,h}^{\text{sink}}; \mathbf{K}_{n,h}^{C_1}; \mathbf{K}_{n,h}^{C_2}; \mathbf{K}_{n,h}^{C_3}]$, where $[\cdot;\cdot]$ denotes concatenation along the key-token dimension. We then compute the attention weights over these blocks:

$$\mathbf{A}_{n,h} = \text{Softmax}\left(\frac{\mathbf{Q}_{n,h}(\mathbf{K}_{n,h}^{\text{cache}})^\top}{\sqrt{d_h}}\right), \tag{6}$$

where d_h denotes the dimension of each attention head. Let $\mathcal{I}_{\text{sink}}$ and $\mathcal{I}_{C_1}$ denote the columns of $\mathbf{A}_{n,h}$ corresponding to $\mathbf{K}_{n,h}^{\text{sink}}$ and $\mathbf{K}_{n,h}^{C_1}$, respectively. We compute their block-level attention masses as

$$s_{n,h}^{\text{sink}} = \frac{1}{N_q}\sum_{q=1}^{N_q}\sum_{j\in\mathcal{I}_{\text{sink}}} \mathbf{A}_{n,h,q,j}, \qquad s_{n,h}^{C_1} = \frac{1}{N_q}\sum_{q=1}^{N_q}\sum_{j\in\mathcal{I}_{C_1}} \mathbf{A}_{n,h,q,j}, \tag{7}$$

where N_q is the number of query tokens. The two scores measure the attention mass assigned to the frame sink and C_1, respectively, averaged over all query tokens. As the earliest retained rolling KV cache block, C_1 serves as a dynamic reference.

We therefore use the reference block as a conservative intervention boundary: regulation is activated only when the frame-sink attention mass exceeds that of C_1, with the regulation strength determined by their ratio. Specifically, we define the adaptive regulation coefficient as follows:

$$\gamma_{n,h} = \begin{cases} \dfrac{s_{n,h}^{C_1}}{s_{n,h}^{\text{sink}} + \epsilon}, & s_{n,h}^{\text{sink}} > s_{n,h}^{C_1}, \\ 1, & s_{n,h}^{\text{sink}} \leq s_{n,h}^{C_1}, \end{cases} \tag{8}$$

where $\epsilon = 10^{-6}$. When $s_{n,h}^{\text{sink}} \leq s_{n,h}^{C_1}$, we set $\gamma_{n,h} = 1$ and leave the corresponding head unchanged. Otherwise, a larger relative frame-sink attention mass yields a smaller regulation coefficient.

The regulation coefficient is applied only to the frame-sink keys of the corresponding head:

$$\widetilde{\mathbf{K}}_{n,h}^{\text{sink}} = \gamma_{n,h} \mathbf{K}_{n,h}^{\text{sink}}. \tag{9}$$

Only the frame-sink keys are modified. All rolling KV cache keys and all cached values remain unchanged, while the original temporal positional encoding is preserved. The regulated frame-sink keys are then used in the subsequent attention computation at the current autoregressive step.

4 Experiments

4.1 Setup

Implementation. We implement CPC on the released LongLive (Yang et al., 2025a) model and apply it entirely at inference time without additional learnable parameters or KV cache expansion. We set $\rho = 0.6$, $\lambda_{\text{init}} = 0.8$, and $\lambda_{\text{cur}} = 1$ for all experiments. Additional implementation and generation details are provided in Appendix C.

Evaluation Protocol. We evaluate CPC on InterVidProM (200 six-stage prompt sequences) at 60- and 120-second horizons and on InterStoryBench (980 multi-stage prompt sequences) at the 60-second horizon. We report five VBench-Long (Huang et al., 2025b) metrics, CLIP Score (Radford et al., 2021), and two reset-oriented metrics: Scene-Reset Rate (SRR) and PSNR-Based Reset-Free Score (PSNR-RFS), which quantify detected returns toward the initial visual state. We compare against eight baselines using identical prompt sequences and transition points. Further details are provided in Appendix C.

4.2 Results in Interactive Long Video Generation

Quantitative results. Table 1 shows that our approach achieves the best subject consistency, background consistency, imaging quality, and aesthetic quality on InterVidProM at both horizons. It also performs strongly on the two reset-oriented metrics, achieving the best SRR and PSNR-RFS at 120 seconds and top-two results at 60 seconds, while maintaining competitive CLIP Scores. On InterStoryBench, our method achieves the best subject consistency, aesthetic quality, SRR, and PSNR-RFS. Overall, CPC improves long-range visual consistency and scene-reset robustness while preserving semantic alignment across different generation settings.

Table 2: User study results. Values are percentages of votes favoring CPC over the baselines.

Method	Sub. Cons. (%) ↑	Instr. Follow. (%) ↑	Temp. Cons. (%) ↑	Overall Qual. (%) ↑
SkyReels-V2	94.6	87.5	94.6	87.5
Self-Forcing	87.7	82.5	91.2	89.5
LongLive	80.7	70.2	78.9	77.2
Rolling Forcing	85.7	80.4	73.2	75.0
Deep Forcing	94.6	78.6	91.1	91.1
Causal Forcing	78.6	75.0	75.0	78.6
LongLive-RAG	78.6	78.6	87.5	80.4
LongLive-LoL	83.9	78.6	82.1	80.4

Qualitative results. Figure 6 illustrates long-range visual consistency across successive prompt stages. Our approach better preserves persistent subject attributes while accommodating evolving actions, viewpoints, and scene content. Baselines exhibit subject drift, visual degradation, or abrupt scene resets. Additional examples are provided in Appendix E.



Figure 6: Qualitative comparison of 60-second interactive video generation. Our approach better preserves long-range visual consistency across prompt stages while adapting to evolving semantic content, whereas baselines exhibit appearance drift, visual degradation, or abrupt scene resets.

User Study. To complement the quantitative and qualitative evaluations, we conduct a 2AFC user study on 60-second videos from InterVid-ProM. Thirty participants each evaluate 15 video pairs for subject consistency, instruction following, temporal consistency, and overall quality. As shown in Table 2, our approach receives a majority of preferences over every baseline across all four criteria. Figure 7 summarizes the macro-averaged preferences across the eight baseline comparisons. The protocol is detailed in Appendix F.

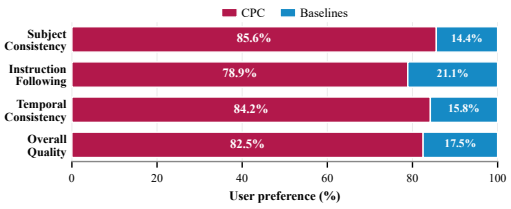

Figure 7: User preference study. CPC shows higher preference across all four criteria.

4.3 Ablation Study

We evaluate the contributions of *Semantic Anchor* and *Adaptive Sink Regulation* on InterVid-ProM using the same LongLive backbone and generation setting. As shown in Table 3, CPC without *Adaptive Sink Regulation* improves VBench-Long Avg. from 84.83 to 84.97, while CPC without *Semantic Anchor* substantially reduces SRR from 32.73 to 0.38 and increases PSNR-

Table 1: Quantitative comparison on InterVidProM at 60- and 120-second horizons and on InterStoryBench at the 60-second horizon. Bold and underlined values indicate the best and second-best results, respectively.

Method	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Imaging Quality ↑	Aesthetic Quality ↑	CLIP Score ↑	SRR ↓ (10^{-4})	PSNR-RFS ↑
InterVidProM 60s generation								
SkyReels-V2	94.64	95.79	**99.15**	52.31	54.80	29.54	0.88	95.89
Self-Forcing	93.94	94.90	97.94	67.10	56.97	28.98	1.51	89.52
LongLive	96.71	95.99	98.73	70.77	61.93	30.97	32.73	48.70
Rolling Forcing	96.75	96.00	98.76	71.07	60.11	29.57	0.31	**99.66**
Deep Forcing	93.55	94.27	97.94	68.06	59.03	30.32	1.76	95.34
Causal Forcing	94.26	94.48	97.38	67.96	55.99	29.38	**0.13**	97.75
LongLive-LoL	95.56	95.60	98.63	70.36	61.51	**31.14**	1.01	98.74
LongLive-RAG	96.48	95.61	98.67	69.93	60.18	30.70	0.82	98.10
CPC (Ours)	**97.09**	**96.29**	98.84	**71.12**	**63.03**	30.88	0.31	99.57
InterVidProM 120s generation								
SkyReels-V2	94.41	95.67	**99.16**	51.61	53.69	29.53	0.28	94.70
Self-Forcing	93.95	95.11	98.04	63.32	53.18	27.41	0.44	88.00
LongLive	96.55	95.69	98.75	70.29	60.79	30.90	27.16	29.95
Rolling Forcing	96.26	95.62	98.75	70.45	58.77	29.48	0.53	97.52
Deep Forcing	92.97	94.08	97.98	68.23	58.43	30.21	1.75	92.93
Causal Forcing	94.45	94.33	97.42	67.46	55.09	29.48	0.41	97.38
LongLive-LoL	96.04	95.63	98.65	69.18	61.35	**31.20**	0.66	98.46
LongLive-RAG	96.73	95.61	98.70	69.53	60.00	30.63	0.19	98.30
CPC (Ours)	**96.97**	**96.03**	98.82	**71.09**	**61.95**	30.69	**0.03**	**99.50**
InterStoryBench 60s generation								
SkyReels-V2	96.03	96.30	**99.23**	65.43	54.20	30.85	2.08	97.64
Self-Forcing	95.96	95.46	98.95	69.34	54.73	30.61	1.44	98.72
LongLive	97.25	95.95	99.04	72.45	55.63	31.96	29.46	58.16
Rolling Forcing	97.51	96.13	98.99	74.11	53.65	30.92	1.55	99.42
Deep Forcing	97.96	**96.53**	99.02	**74.36**	55.60	31.54	14.68	77.35
Causal Forcing	94.72	94.49	97.83	70.72	50.46	30.22	0.22	99.64
LongLive-LoL	97.55	96.02	99.06	72.84	55.11	**32.05**	0.54	99.13
LongLive-RAG	97.24	95.76	98.92	72.09	54.66	31.88	1.23	97.04
CPC (Ours)	**98.09**	96.43	99.19	72.79	**55.77**	31.66	**0.12**	**99.96**

RFS from 48.70 to 99.69 with no change in VBench-Long Avg. Combining both components, CPC achieves the highest VBench-Long Avg. of 85.27 and the lowest SRR of 0.31, while maintaining a PSNR-RFS of 99.57. CLIP Scores remain similar across all variants. These results indicate complementary roles of *Semantic Anchor* and *Adaptive Sink Regulation* in improving overall generation quality and suppressing scene resets. Additional qualitative ablations are provided in Appendix G.

Table 3: Ablation on InterVidProM. VBench-Long Avg. is the five-dimension average.

Method	VBench-Long Avg. ↑	CLIP Score ↑	SRR ↓ (10^{-4})	PSNR-RFS ↑
LongLive	84.83	30.97	32.73	48.70
CPC w/o Adaptive Sink Regulation	84.97	30.70	34.80	46.42
CPC w/o Semantic Anchor	84.83	**31.05**	0.38	**99.69**
CPC	**85.27**	30.88	**0.31**	99.57

5 Conclusion

In this paper, we addressed the insufficient persistence of textual conditioning in interactive long video generation. We proposed Context Persistence Calibration (CPC), a training-free framework that maintains access to the initial textual context while accommodating evolving instructions. *Semantic Anchor* jointly contextualizes the initial and current prompts and retains the leading singular components of the contextualized initial-prompt representation to construct a rank-constrained textual reference. To further mitigate *sink-collapse*, *Adaptive Sink Regulation* selectively regulates excessive frame-sink contributions. Experiments on InterVidProM and InterStoryBench demonstrate improved long-range visual consistency, fewer scene resets, and continued responsiveness to evolving prompts. Limitations are discussed in Appendix B.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist with the retrieval and discovery of related work, and generate synthetic prompt data. All AI-assisted writing was reviewed by the authors, and all identified references were independently verified. We additionally used generative AI to construct multi-stage prompt sequences for InterVidProM and InterStoryBench by expanding prompts derived from VidProM and StoryBench. The construction procedure for the evaluation sets and corresponding instruction templates are described in the paper and Appendix C.2. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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.

Emanuele Bugliarello, H Hernan Moraldo, Ruben Villegas, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Han Zhang, Dumitru Erhan, Vittorio Ferrari, Pieter-Jan Kindermans, and Paul Voigtlaender. Storybench: A multifaceted benchmark for continuous story visualization. *Advances in Neural Information Processing Systems*, 36:78095–78125, 2023.

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. *arXiv preprint arXiv:2504.13074*, 2025.

Jiayu Chen, Junbei Tang, Wenbiao Zhao, Maoliang Li, Jiayi Luo, Zihao Zheng, Jiawei Yang, Guojie Luo, and Xiang Chen. Pyramid forcing: Head-aware pyramid kv cache policy for high-quality long video generation. *arXiv preprint arXiv:2605.13111*, 2026a.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Zhou Daquan, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. In *International Conference on Learning Representations*, volume 2026, pp. 38973–38997, 2026b.

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*, 2026c.

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.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Lol: Longer than longer, scaling video generation to hour. *arXiv preprint arXiv:2601.16914*, 2026.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European conference on computer vision*, pp. 102–118. Springer, 2022.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. Longlive-rag: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2025a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

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.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026a.

Jia Li, Xiaomeng Fu, Xurui Peng, Weifeng Chen, Youwei Zheng, Tianyu Zhao, Jiexi Wang, Fangmin Chen, Xing Wang, and Hayden Kwok-Hay So. Train short, inference long: Training-free horizon extension for autoregressive video generation. *arXiv preprint arXiv:2602.14027*, 2026b.

Zhuoling Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Longdiff: Training-free long video generation in one go. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17789–17798. IEEE, 2025.

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*, 2025.

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, 2025.

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.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026. URL `https://qwen.ai/blog?id=qwen3.6`.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

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.

Jing Wang, Xiangxin Zhou, Jiajun Liang, Kaiqi Liu, Wanyuan Pang, Zhenyu Xie, Tianyu Pang, and Xiaodan Liang. Tempact: Advancing temporal plausibility in autoregressive video generation via planner-executor rl, 2026. URL https://arxiv.org/abs/2606.28016.

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

Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, et al. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

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.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *International Conference on Learning Representations*, volume 2024, pp. 765–783, 2024.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, et al. Videossm: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626v1*, 2025. URL https://arxiv.org/abs/2504.12626v1.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 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.

Algorithm 1 Interactive Inference with Context Persistence Calibration

Require: Frozen causal video generator G_θ and T5 text encoder T_ϕ
Require: Prompt sequence $\mathcal{P} = [P_1, \dots, P_M]$; switch indices $\mathcal{S} = [s_2, \dots, s_M]$
Require: Number of video chunks N; denoising timesteps $t_T > \cdots > t_1 > 0$; re-noising operator Ψ
Require: Retention ratio ρ; weights $\lambda_{\text{init}}, \lambda_{\text{cur}}$; regulation constant ϵ

Initialize output $\mathcal{X} \leftarrow []$ and KV cache $\mathcal{C} \leftarrow \emptyset$
$i \leftarrow 1$; $\mathbf{E}^{\text{act}} \leftarrow T_\phi(P_1)$
for $n = 1, \dots, N$ **do**
  if $n \in \mathcal{S}$ **then**
    $i \leftarrow i + 1$
    $\widetilde{\mathbf{E}}_i \leftarrow T_\phi(P_1 \oplus P_i)$
    Identify the initial and current token regions $\mathcal{I}_i^{\text{init}}$ and $\mathcal{I}_i^{\text{cur}}$
    $\widetilde{\mathbf{E}}_i^{\text{init}} \leftarrow \widetilde{\mathbf{E}}_i[\mathcal{I}_i^{\text{init}}, :]$ // Eq. 2
    $(L, d) \leftarrow \text{shape}(\widetilde{\mathbf{E}}_i^{\text{init}})$; $r \leftarrow \min(L, d)$
    $(\mathbf{U}_i, \boldsymbol{\sigma}_i, \mathbf{V}_i) \leftarrow \text{SVD}(\widetilde{\mathbf{E}}_i^{\text{init}})$ // Eq. 3
    $k \leftarrow \max(1, \lfloor \rho r \rfloor)$
    $\widehat{\mathbf{E}}_i^{\text{init}} \leftarrow \mathbf{U}_i[:, 1:k]\ \text{diag}(\sigma_{i,1}, \dots, \sigma_{i,k})\ \mathbf{V}_i[:, 1:k]^\top$ // Eq. 4
    $\widehat{\mathbf{E}}_i \leftarrow \widetilde{\mathbf{E}}_i$
    $\widehat{\mathbf{E}}_i[\mathcal{I}_i^{\text{init}}, :] \leftarrow \lambda_{\text{init}} \widehat{\mathbf{E}}_i^{\text{init}}$ // Eq. 5
    $\widehat{\mathbf{E}}_i[\mathcal{I}_i^{\text{cur}}, :] \leftarrow \lambda_{\text{cur}} \widetilde{\mathbf{E}}_i[\mathcal{I}_i^{\text{cur}}, :]$ // Eq. 5
    $\mathbf{E}^{\text{act}} \leftarrow \widehat{\mathbf{E}}_i$
    $\mathcal{C} \leftarrow \text{Recache}(G_\theta, \mathcal{X}, \mathcal{C}, \mathbf{E}^{\text{act}})$
  end if
  Sample $x_{n,t_T} \sim \mathcal{N}(0, \mathbf{I})$
  for $j = T, \dots, 1$ **do**
    $\widehat{x}_{n,0} \leftarrow G_\theta^{\text{ASR}}(x_{n,t_j}; t_j, \mathcal{C}, \mathbf{E}^{\text{act}})$ // Eqs. 6–9
    if $j = 1$ **then**
      Append $\widehat{x}_{n,0}$ to $\mathcal{X}$
      $\mathcal{C} \leftarrow G_\theta^{\text{KV}}(\widehat{x}_{n,0}, 0, \mathcal{C}, \mathbf{E}^{\text{act}})$
    else
      Sample $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I})$
      $x_{n,t_{j-1}} \leftarrow \Psi(\widehat{x}_{n,0}, \boldsymbol{\eta}, t_{j-1})$
    end if
  end for
end for
return $\mathcal{X}$

A Algorithm Detail of Context Persistence Calibration

Algorithm 1 summarizes the interactive inference procedure of CPC, following the prompt-switching and KV-recache workflow of LongLive (Yang et al., 2025a). Here, N denotes the number of video chunks, and s_i denotes the first chunk generated under prompt P_i, with $1 < s_2 < \cdots < s_M \leq N$. The first stage uses $T_\phi(P_1)$ without modification. At each prompt transition, *Semantic Anchor* constructs the calibrated textual condition from P_1 and the newly received prompt P_i.

We denote a generator forward pass incorporating *Adaptive Sink Regulation* by G_θ^{ASR}. Within the corresponding attention computations, frame-sink keys are regulated using the current queries according to Eqs. 6–9, while rolling KV cache keys and all cached values remain unchanged. Recache refreshes the rolling KV cache while preserving the frame sink, and G_θ^{KV} updates the cache using the generated clean chunk under the original bounded-cache policy. Ψ denotes the re-noising operation used by the base sampler.

B Limitation Analysis

While CPC improves long-range consistency by maintaining a persistent textual reference to the initial prompt, it does not explicitly establish new long-term anchors for attributes introduced only in later interaction stages. If a new subject, object, or visual attribute appears in a subsequent prompt and is expected to persist across multiple later stages, its consistency must still be maintained largely through the evolving visual context and current instructions. As a result, such newly introduced

attributes may gradually drift over extended interactions. Extending CPC to dynamically promote persistent information introduced at later stages into the textual reference is an important direction for future work.

C Implementation Details

C.1 Generation and CPC Settings

We follow the LongLive (Yang et al., 2025a) generation configuration, with one persistent frame sink, three rolling KV cache blocks, and the original KV recache procedure. All model parameters remain frozen. The first stage uses the initial prompt without modification; subsequent stages use the textual condition constructed by *Semantic Anchor*. All videos are generated at 16 FPS and a resolution of 832×480. All experiments are conducted on NVIDIA A40 GPUs.

For a video of duration D with M interaction stages, each prompt controls a segment of duration D/M. InterVidProM therefore uses six 10-second segments for 60-second generation and six 20-second segments for 120-second generation. InterStoryBench uses 60-second videos, with segment durations determined by the number of prompts in each sequence.

C.2 Evaluation Set Construction

For InterVidProM, we randomly sample 200 source prompts from VidProM (Wang & Yang, 2024) and expand each into a six-stage prompt sequence using Qwen3.6-Plus (Qwen Team, 2026). The first stage remains visually close to the source prompt, while later stages develop through coherent actions, events, and scene changes. For InterStoryBench, we retain 980 StoryBench (Bugliarello et al., 2023) sequences with at least three prompt segments and use the same model to enrich each segment with visual details while preserving the original semantics, narrative order, and number of stages. The instruction templates used for the two evaluation sets are provided below.

InterVidProM.

```
You are a video-prompt generation specialist. Your task:
• Receive an ORIGINAL_PROMPT as the starting point of a continuous video.
• Generate exactly six English text-to-video prompts forming one coherent
six-stage interaction sequence.
• Keep the first prompt visually close to the ORIGINAL_PROMPT, while
allowing later prompts to naturally develop the scene through new
actions, events, interactions, environmental changes, or camera movements.
• Maintain visual and narrative continuity between adjacent prompts and
keep the main subject recognizable unless the story naturally shifts focus.
• Introduce a meaningful new action, interaction, object, environmental
change, camera movement, or story event in each segment. Avoid repetitive
actions and abrupt unrelated scene changes.
• Write each prompt as a detailed cinematic description suitable for
text-to-video generation.
• Do not include explanations, numbering, timestamps, or additional text.
```

InterStoryBench.

```
You are a video-prompt enrichment specialist. Your task:
• Receive a multi-stage prompt sequence describing one continuous video.
• Enrich each prompt with concrete visual details while preserving the
original semantics, narrative structure, and number of stages.
• Preserve the original order, subjects, actions, settings, objects, and
events without introducing new story developments.
• Maintain visual and temporal continuity between adjacent prompts.
• Add appropriate details about appearance, environment, lighting, motion,
composition, and camera framing or movement.
• If a prompt is already sufficiently detailed, refine it only lightly.
• Return the same number of prompts as the input without explanations,
numbering, timestamps, or additional text.
```

C.3 Baseline Configurations

We compare CPC with LongLive (Yang et al., 2025a), Self-Forcing (Huang et al., 2025a), SkyReels-V2 (Chen et al., 2025), Rolling Forcing (Liu et al., 2025), Deep Forcing (Yi et al., 2025), Causal Forcing (Zhu et al., 2026), LongLive-RAG (Hu et al., 2026), and LongLive-LoL (Cui et al., 2026). We use officially released checkpoints and original inference configurations whenever available. As no official implementation of LoL is available, we implement it on the LongLive backbone following the procedure described in the paper and denote this baseline as LongLive-LoL.

LongLive natively supports the multi-stage interactive long-video setting considered in our evaluation, and our LongLive-LoL implementation inherits this capability from the LongLive backbone. For the remaining baselines, whose original inference pipelines do not directly support repeated prompt switches over long generation horizons, we update the active textual condition at predefined transition points while continuing autoregressive generation. These adaptations require no additional training or architectural modifications. All methods are evaluated using identical prompt sequences and transition points.

C.4 Evaluation Metrics

We use the official VBench-Long evaluation pipeline to compute Subject Consistency, Background Consistency, Motion Smoothness, Imaging Quality, and Aesthetic Quality. For CLIP Score, we evaluate each video segment against its corresponding prompt and average the stage-level scores to obtain the video-level score.

To evaluate scene-reset behavior, we report two complementary metrics: Scene-Reset Rate (SRR) and PSNR-Based Reset-Free Score (PSNR-RFS). Scene-Reset Rate (SRR) measures the proportion of evaluated frame transitions satisfying the reset criterion, while PSNR-Based Reset-Free Score (PSNR-RFS) is derived from the accumulated magnitude of positive PSNR jumps toward the initial frame. Larger or more frequent positive jumps increase the accumulated penalty and therefore decrease PSNR-RFS. Accordingly, lower SRR and higher PSNR-RFS indicate fewer or weaker detected returns toward the initial visual state. For readability, we report $10^4 \times \mathrm{SRR}$, corresponding to detected reset transitions per 10,000 evaluated frame transitions, while PSNR-RFS is reported on a 0–100 scale. Both reset metrics are macro-averaged over videos. Detailed definitions of SRR and PSNR-RFS are provided in Sec. C.5. In Table 3, VBench-Long Avg. is the arithmetic mean of the five VBench-Long dimensions.

C.5 Scene-Reset Evaluation

We use two complementary pixel-level metrics to approximately quantify scene-reset behavior: Scene-Reset Rate (SRR) and PSNR-Based Reset-Free Score (PSNR-RFS). Both use the first generated frame as a fixed visual reference and examine changes in similarity to this reference. To focus on returns occurring after the initial interaction stage, we exclude the entire first stage, during which high similarity to the initial frame may naturally arise under the same prompt.

For video v, let $x_{v,1}$ denote its first generated frame, F_v its total number of frames, and K_v its number of successive prompts. Under the equal-duration stage setting, the number of frames excluded from the beginning of the video is

$$B_v = \mathrm{round}\left(\frac{F_v}{K_v}\right). \tag{10}$$

The ordered indices of the evaluated frames are $\mathcal{T}_v = \{B_v + 1, \ldots, F_v\} = \{t_1, \ldots, t_{N_v}\}$, where $N_v = F_v - B_v$. Each evaluated frame is resized to match the initial frame when necessary. Similarity changes are computed only between adjacent evaluated frames. Thus, the transition from the final frame of the first stage to the first frame of the second stage is not evaluated. The definitions below assume $N_v \geq 2$.

Scene-Reset Rate. For each evaluated frame x_{v,t_m}, we compute its pixel-space cosine similarity to the initial frame:

$$c_{v,m} = \frac{\mathrm{vec}(x_{v,t_m})^\top \mathrm{vec}(x_{v,1})}{\|\mathrm{vec}(x_{v,t_m})\|_2 \|\mathrm{vec}(x_{v,1})\|_2}, \qquad m = 1, \ldots, N_v, \tag{11}$$

where $\mathrm{vec}(\cdot)$ flattens a frame into a vector.

High similarity alone does not necessarily indicate a scene reset, since persistent visual content may remain similar to the initial frame. We therefore flag an evaluated frame transition only when the current similarity exceeds an absolute threshold and increases sufficiently relative to the preceding evaluated frame:

$$r_{v,m} = \mathbb{I}[c_{v,m} > \tau_{\text{high}} \ \wedge \ c_{v,m} - c_{v,m-1} > \tau_{\text{jump}}], \qquad m = 2, \ldots, N_v, \tag{12}$$

where $\mathbb{I}[\cdot]$ is the indicator function. We use $\tau_{\text{high}} = 0.85$ and $\tau_{\text{jump}} = 0.02$ for all compared methods.

The video-level SRR is defined as the proportion of evaluated frame transitions satisfying this criterion:

$$\mathrm{SRR}_v = \frac{\sum_{m=2}^{N_v} r_{v,m}}{N_v - 1}. \tag{13}$$

We compute the dataset-level SRR by averaging these proportions equally over videos:

$$\mathrm{SRR} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathrm{SRR}_v, \tag{14}$$

where $\mathcal{V}$ denotes the evaluation set. A lower SRR indicates a smaller proportion of evaluated frame transitions satisfying the reset criterion. Consecutive flagged transitions are counted separately rather than merged into distinct reset episodes.

For readability, all tables report SRR in units of 10^{-4}, i.e., each displayed value corresponds to $10^4\,\mathrm{SRR}$ and can be interpreted as the macro-averaged number of detected reset transitions per 10,000 evaluated frame transitions.

PSNR-Based Reset-Free Score. Whereas SRR counts transitions satisfying the cosine-similarity criterion, PSNR-RFS measures the accumulated magnitude of sufficiently large positive PSNR changes relative to the initial frame.

For frames x and y, we compute the mean squared error as

$$\mathrm{MSE}(x, y) = \frac{1}{HWC} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} (x_{i,j,k} - y_{i,j,k})^2, \tag{15}$$

where H, W, and C denote the height, width, and number of channels of the evaluated image arrays. Let $e_{v,m} = \mathrm{MSE}(x_{v,t_m}, x_{v,1})$. For 8-bit images, the raw PSNR is

$$\widetilde{p}_{v,m} = \begin{cases} 10 \log_{10}\left(\dfrac{L^2}{e_{v,m}}\right), & e_{v,m} > 0, \\ +\infty, & e_{v,m} = 0, \end{cases} \qquad L = 255. \tag{16}$$

Following the implementation, finite PSNR values are capped at $P_{\max} = 100$ dB, and non-finite values are replaced by this cap:

$$p_{v,m} = \begin{cases} \min(\widetilde{p}_{v,m}, P_{\max}), & \widetilde{p}_{v,m} \text{ is finite}, \\ P_{\max}, & \text{otherwise}. \end{cases} \tag{17}$$

All subsequent computations use these capped PSNR values.

Positive PSNR changes indicate increased pixel-level similarity to the reference frame. We compute

$$\Delta p_{v,m} = p_{v,m} - p_{v,m-1}, \qquad m = 2, \ldots, N_v, \tag{18}$$

and retain only changes exceeding a fixed positive threshold:

$$q_{v,m} = \mathbb{I}[\Delta p_{v,m} > \tau_{\text{PSNR}}], \qquad m = 2, \ldots, N_v, \tag{19}$$

where $\tau_{\text{PSNR}} = 1$ dB for all compared methods. We do not impose an additional threshold on the absolute PSNR after the jump. Negative changes and positive changes not exceeding 1 dB are not penalized.

We accumulate the full magnitudes of the qualifying positive jumps:

$$A_v = \sum_{m=2}^{N_v} q_{v,m} \Delta p_{v,m}. \tag{20}$$

The accumulated magnitude is normalized by the mean PSNR over the evaluated frames:

$$\bar{p}_v = \frac{1}{N_v} \sum_{m=1}^{N_v} p_{v,m}. \tag{21}$$

The implementation requires $\bar{p}_v$ to be finite and positive. The video-level reset-free score is then

$$S_v^{\mathrm{PSNR}} = \exp\left(-\frac{A_v}{\bar{p}_v}\right). \tag{22}$$

The normalized quantity $A_v/\bar{p}_v$ is a non-negative penalty. The negative exponential converts a larger penalty into a lower score, so higher values indicate smaller normalized positive-jump penalties. For finite A_v and positive $\bar{p}_v$, $S_v^{\mathrm{PSNR}} \in (0, 1]$. If no positive jump exceeds the threshold, then $A_v = 0$ and $S_v^{\mathrm{PSNR}} = 1$.

We report PSNR-RFS as the macro-average of the video-level scores, rescaled to 0–100:

$$\text{PSNR-RFS} = \frac{100}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} S_v^{\mathrm{PSNR}}. \tag{23}$$

Thus, the exponential transformation is applied separately to each video before averaging. The factor of 100 changes only the reporting scale. For fewer than two evaluated frames, the implementation returns $\mathrm{SRR}_v = 0$ and $S_v^{\mathrm{PSNR}} = 1$ by convention.

Together, SRR and PSNR-RFS provide complementary pixel-level approximations of scene-reset behavior: SRR measures the proportion of transitions satisfying the reset criterion, while PSNR-RFS reflects a normalized penalty for the accumulated magnitude of positive PSNR jumps. Lower SRR and higher PSNR-RFS are preferable. Neither metric directly measures how long a reset state persists. Since PSNR-RFS accumulates jump magnitudes without normalizing by the number of transitions, it is compared within matched generation settings rather than treated as invariant to video duration. We analyze the sensitivity of both metrics to their threshold parameters in Sec. C.6.

C.6 Parameter Sensitivity of Scene-Reset Metrics

We further evaluate the sensitivity of the scene-reset metrics to their threshold parameters under the 60-second InterVidProM setting. For SRR, we vary $\tau_{\mathrm{high}} \in [0.75, 0.95]$ and $\tau_{\mathrm{jump}} \in [0.005, 0.05]$. As shown in Fig. 8(a), method rankings remain stable around the default setting $(\tau_{\mathrm{high}}, \tau_{\mathrm{jump}}) = (0.85, 0.02)$. The eight neighboring configurations yield a mean Kendall's τ_b of 0.84 relative to the default ranking, indicating robustness to moderate perturbations of either threshold. Ranking stability decreases under some overly stringent threshold settings, whereas the default lies within a broad stable region.

For PSNR-RFS, we vary $\tau_{\mathrm{PSNR}} \in [0.25, 3.0]$ dB. Fig. 8(b) shows that increasing the threshold progressively drives many methods toward the upper end of the score range, reducing discriminability. Nevertheless, as shown in Fig. 8(c), method rankings remain highly consistent for $\tau_{\mathrm{PSNR}} \in [0.75, 1.5]$ dB, with Kendall's τ_b ranging from 0.89 to 1.00. The default $\tau_{\mathrm{PSNR}} = 1$ dB lies within this stable region while avoiding the stronger score saturation observed at larger thresholds. Overall, the default thresholds lie in stable, non-degenerate operating regions, and the comparative conclusions are robust to moderate parameter variations.

D Qualitative Analysis of Direct Prompt Concatenation

We examine whether directly retaining the initial prompt is sufficient for interactive long video generation. At each subsequent stage, direct prompt concatenation combines the initial prompt P_1 with the current prompt P_i, retaining the complete initial instruction without singular-value truncation.

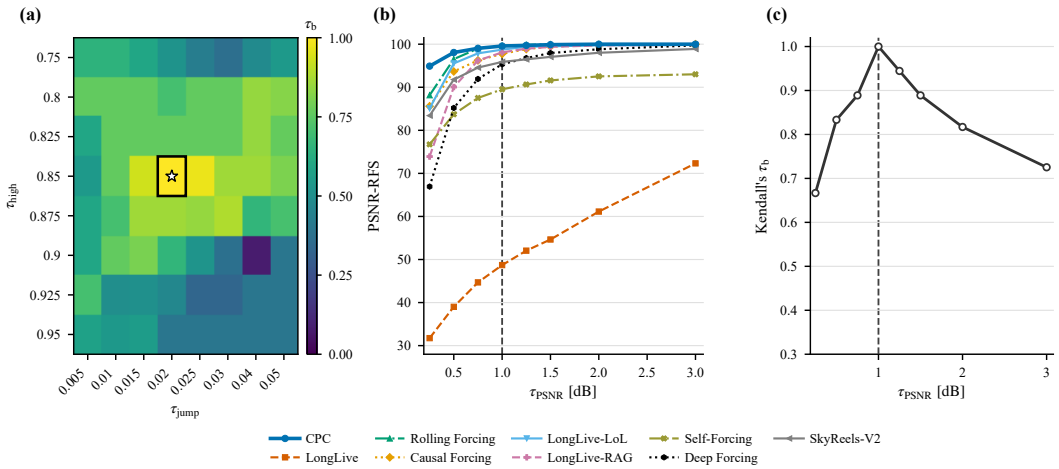

Figure 8: Parameter sensitivity of the scene-reset metrics on InterVidProM. (a) SRR ranking stability under different $(\tau_{\text{high}}, \tau_{\text{jump}})$ settings, measured by Kendall's τ_b against the default ranking. (b) PSNR-RFS scores under different τ_{PSNR} values. (c) PSNR-RFS ranking stability relative to the default setting. The star in (a) and dashed lines in (b)–(c) indicate the default parameters.

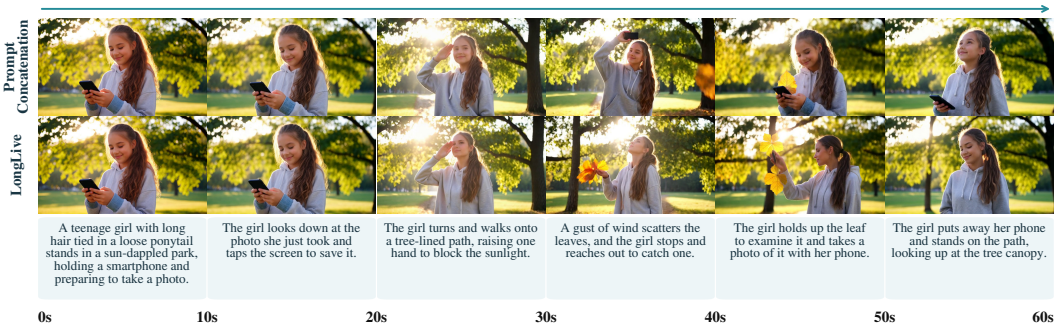

Figure 9: Qualitative comparison of direct prompt concatenation and LongLive. Direct prompt concatenation shows weakened adherence to later instructions, particularly in the 30–60 s stages.

Figures 9 and 10 compare this strategy with LongLive and illustrate how outdated stage-specific content can interfere with evolving instructions.

In Fig. 9, the initial prompt describes a girl holding a smartphone and preparing to take a photo, whereas later prompts introduce interactions with leaves and eventually require her to put the phone away. With direct prompt concatenation, the sampled frame at 30–40 s shows the girl raising her phone above her head rather than reaching out to catch a leaf. At 40–50 s, a yellow leaf appears near the phone, but the requested action of holding it up for inspection is not clearly depicted. At 50–60 s, she looks upward but still holds the phone, only partially following the instruction to put it away and look at the tree canopy. These frames illustrate incomplete adherence to later instructions while the initial phone-related activity remains prominent.

Figure 10 shows a related failure involving the persistence and recurrence of early-stage actions. The green dragon initially eats spaghetti, but subsequent prompts require it to push the plate aside, move onto the dance floor, and perform a final arms-raised pose. Under direct prompt concatenation, eating-related content persists at 20–40 s despite the requested transition away from the initial activity. Although the 40–50 s frame shows the dragon on the dance floor, the final frame returns to a table-and-spaghetti composition rather than clearly depicting the requested pose. Together, these examples illustrate that directly retaining the full initial instruction can interfere with later actions, motivating calibration of the initial textual context rather than its unmodified reuse.

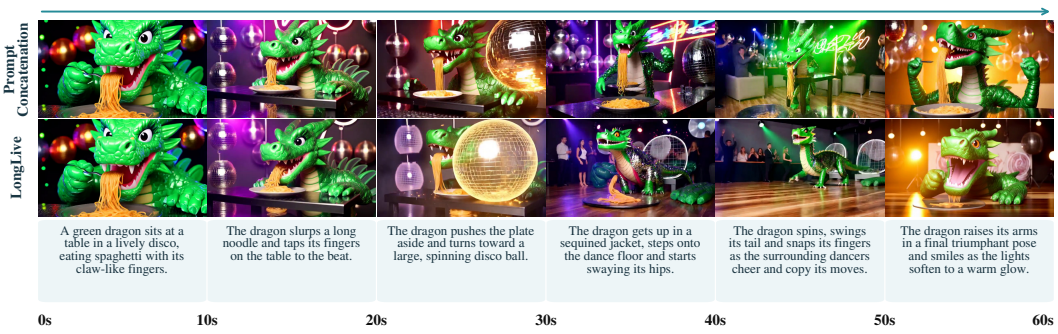

Figure 10: Qualitative comparison of direct prompt concatenation and LongLive. Direct prompt concatenation causes the initial spaghetti-eating action to persist or reappear, interfering with the subsequent dance instructions.

E Additional Qualitative Results

Figures 11 and 12 provide additional CPC results on InterVidProM at 60-second and 120-second generation horizons, respectively. Across successive prompt transitions, CPC preserves persistent subject and scene attributes while following changes in actions, viewpoints, and newly introduced objects.

Figure 13 provides an additional 60-second CPC result on InterStoryBench. Across successive prompt transitions, CPC preserves the subject and scene attributes while following the specified changes in pose and motion.

F User Study Protocol

We conduct human evaluation on InterVidProM following a two-alternative forced-choice (2AFC) protocol. We select 15 interactive prompt sequences for the study. For each prompt sequence, we generate nine videos, one using our approach and eight using the compared baselines. We then pair the video generated by our approach with each baseline result from the same prompt sequence, resulting in 120 distinct candidate video pairs in total.

Thirty participants each evaluate one questionnaire containing 15 video pairs randomly sampled from this candidate pool. We approximately balance the assignment of pairs involving different baselines across participants to avoid overrepresenting any particular baseline. We ensure that the same video pair does not appear more than once within a questionnaire. The generating methods are anonymized as Model A and Model B and are not disclosed to the participants.

For every video pair, participants select the preferred video according to four evaluation criteria: (1) Subject Consistency—Which video keeps the main subject more visually consistent throughout the video? (2) Instruction Following—Which video follows the given prompt or instructions more accurately? (3) Temporal Consistency—Which video is more temporally consistent, with smoother motion and fewer sudden visual changes? (4) Overall Quality—Overall, which video has better visual quality and appears more natural?

The study yields 450 pairwise evaluations and 1,800 criterion-level binary decisions. For each baseline and criterion, Table 2 reports the percentage of valid decisions favoring CPC among comparisons against that baseline. Figure 7 further summarizes the aggregate preferences across the eight baseline comparisons. CPC achieves aggregate preference rates of 85.4%, 79.0%, 84.2%, and 82.4% for subject consistency, instruction following, temporal consistency, and overall quality, respectively.

G Qualitative Ablation Study

Figure 14 qualitatively illustrates the contributions of *Semantic Anchor* and *Adaptive Sink Regulation* in a representative sequence. LongLive exhibits both gradual appearance drift and abrupt scene

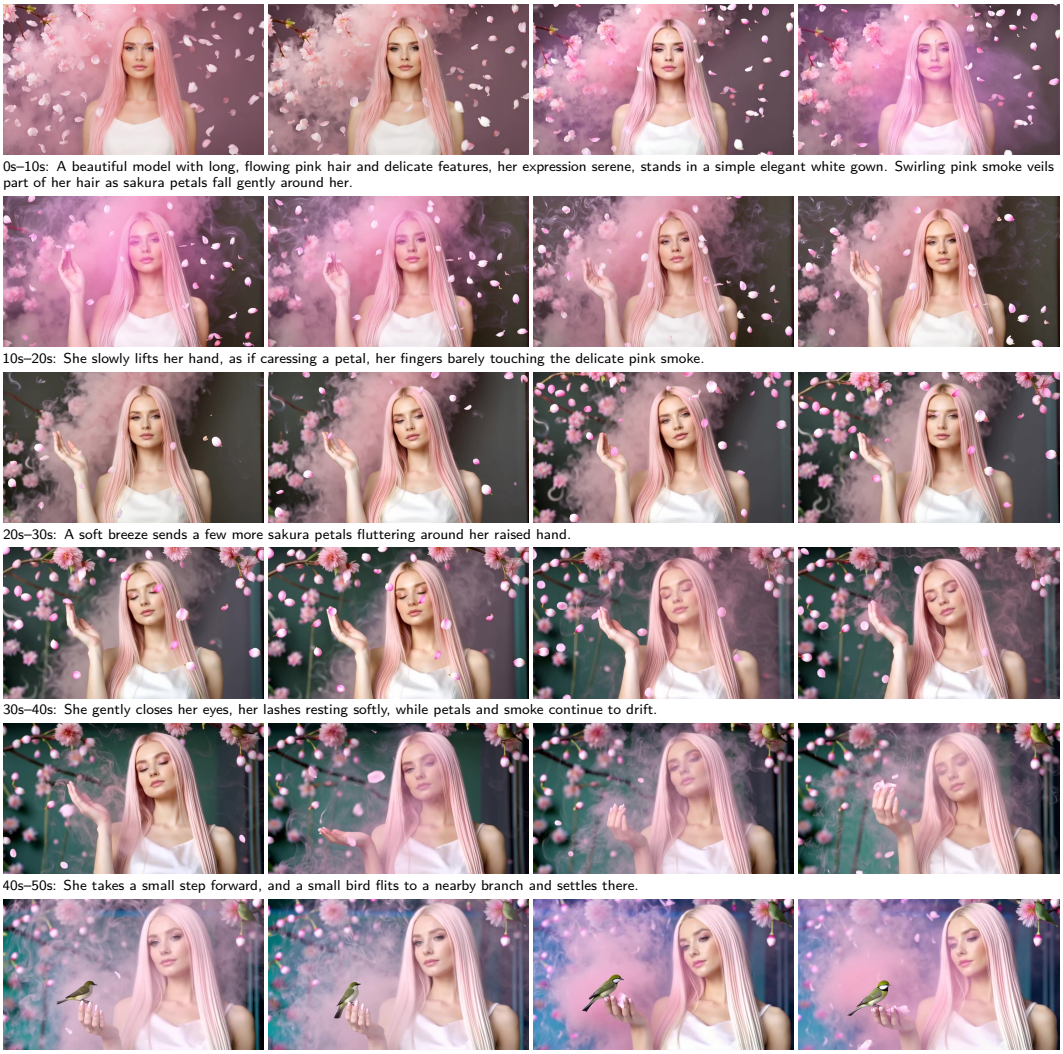

Figure 11: Additional 60-second CPC result on InterVidProM. Each row corresponds to one of six successive 10-second prompt segments, with four frames sampled within each segment. CPC consistently preserves the subject's pink hair, white clothing, and surrounding cherry-blossom scene while following the evolving actions and the introduction of a bird.

resets toward earlier visual states. CPC without *Adaptive Sink Regulation* retains the textual anchor, helping preserve persistent subject appearance, but remains susceptible to abrupt scene resets. In contrast, CPC without *Semantic Anchor* suppresses such resets through sink regulation, while gradual appearance drift may still accumulate without persistent textual guidance. Combining both components, CPC better preserves persistent subject attributes while avoiding regression toward earlier visual states.

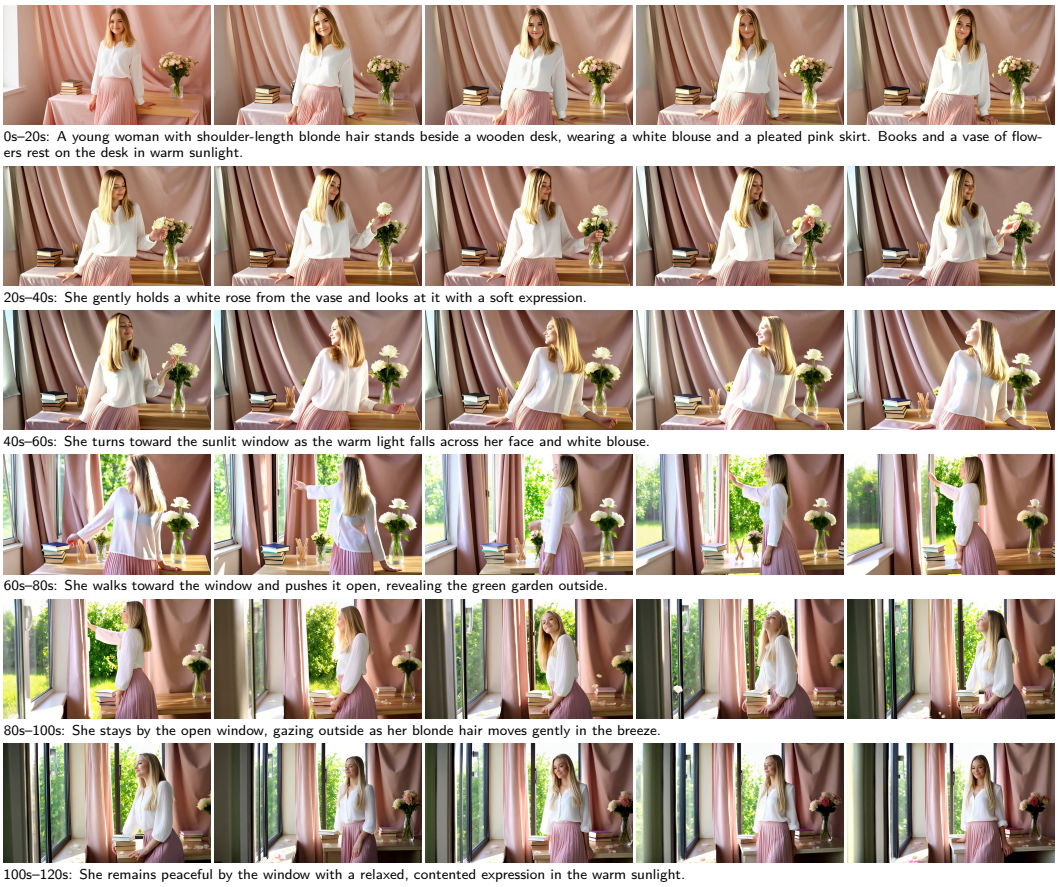

Figure 12: Additional 120-second CPC result on InterVidProM. Each row corresponds to one of six successive 20-second prompt segments, with five frames sampled within each segment. CPC maintains the subject identity, clothing, and indoor scene across the extended generation while following the transition from the desk to the open window and the outdoor garden.

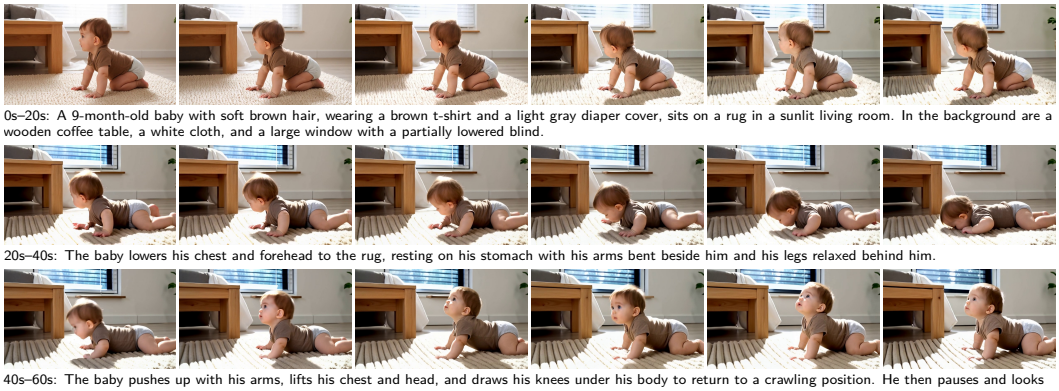

Figure 13: Additional 60-second CPC result on InterStoryBench. Each row corresponds to one of three successive 20-second prompt segments, with six frames sampled within each segment. CPC preserves the baby's appearance, clothing, and living-room environment while following the specified pose and motion transitions.

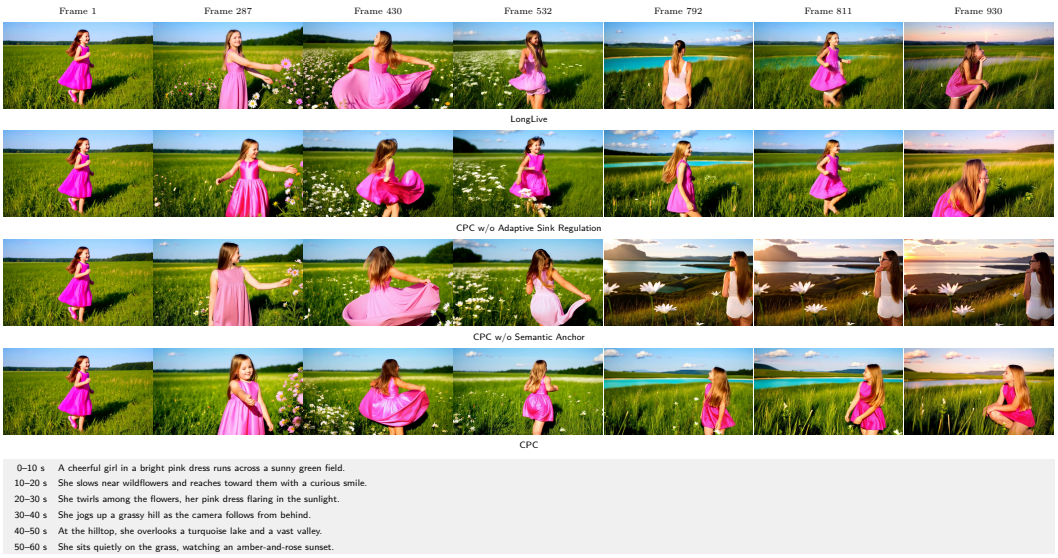

Figure 14: Qualitative ablation of CPC on InterVidProM. CPC w/o *Adaptive Sink Regulation* better preserves persistent appearance attributes but remains susceptible to scene resets, while CPC w/o *Semantic Anchor* suppresses such resets but may still exhibit gradual appearance drift. Full CPC combines both effects for improved long-range visual consistency.