SUBJECTSWITCH: BACKGROUND-PRESERVING SUBJECT SWITCHING IN INTERACTIVE LONG VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When a new text prompt requests a subject change during interactive long video generation, the video generation model must replace the existing subject while preserving background context that remains compatible with the new instruction. Yet, existing autoregressive models struggle with this balance, since they commonly use a key–value (KV) cache to reuse representations of previously generated frames. To address these limitations, we introduce SubjectSwitch, a training-free framework that selectively updates the recent KV cache and controls the influence of the persistent attention sink. For the recent KV cache, SubjectSwitch first identifies subject changes between consecutive prompts. It then suppresses background responses in subject attention by subtracting background attention, thereby locating the subject-related regions of the KV cache. These regions are selectively reconstructed under the new prompt, while the background regions of the KV cache are preserved. However, selectively updating the recent cache alone is insufficient to prevent subject leakage; the persistent attention sink still continuously broadcasts visual cues from the initial frames, causing the obsolete subject to reappear later in the video. To resolve this, we estimate its influence at selected generation blocks and denoising timesteps and selectively correct predictions in the affected regions, reducing the reappearance of initial-frame content in later frames. We evaluate 60-second generation with six sequential prompts on VidProM-Switch and MotionBind-Switch. SubjectSwitch achieves the best results among the compared methods on Subject Switch Cleanliness, Clean Switch Rate, and Switch Latency Score, improving Clean Switch Rate over LongLive, the strongest baseline on this metric, by 23.43 and 26.60 percentage points, respectively, while keeping the five reported VBench-Long scores within 1.3 points of LongLive.



Figure 1: **Subject switching with SubjectSwitch.** Subject changes follow prompt updates while maintaining background continuity in forest (top) and clay-court (bottom) settings.

1 INTRODUCTION

Recent progress in autoregressive video generation, exemplified by real-time methods such as Causal Forcing (Zhu et al., 2026), has enabled increasingly responsive interaction during video synthesis. Interactive long-video generation allows users to guide an extended video through successive text

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.

prompts, updating their instructions as generation proceeds to control the content of subsequent frames. These updates can introduce or replace subjects, modify their actions, or change the background, allowing users to progressively shape the video. The goal is to incorporate these evolving instructions into a continuous video while maintaining coherent motion and visual continuity across prompt changes.

Among existing autoregressive methods, causal models such as CausVid (Yin et al., 2025) and Self-Forcing (Huang et al., 2025a) support efficient streaming video generation through key–value (KV) caching. LongLive (Yang et al., 2026) combines a rolling cache of recent frames with a persistent frame sink to maintain local continuity and long-range visual context. Upon prompt changes, its KV-recache mechanism recomputes cached representations from historical frames under the new prompt. Grounded Forcing (Chen et al., 2026) further separates global semantic memory from local dynamics and adjusts cache update strengths according to temporal proximity. Echo-Forcing (Wu et al., 2026) organizes historical states into hierarchical memories and performs spatially selective forgetting based on feature differences between historical tokens and a new-scene reference. These methods manage cached information through retention, refresh, and selective forgetting to balance prompt adaptation with visual continuity. However, these mechanisms may not adequately capture the distinct update requirements of subjects and backgrounds during interactive prompt changes: a new instruction may require changing the subject while preserving the background, or changing the background while preserving the subject. As a result, cached content that should change may persist, while context that should remain consistent may be disrupted. Meanwhile, in methods that retain a persistent attention sink, the sink can continue to supply visual cues from the initial frames even after prompt changes, potentially causing content from those frames to reappear in later frames.

Subject-centric prompt transitions require updating subject-related representations while preserving reusable background context. Uniform cache re-encoding does not explicitly distinguish these regions and may retain appearance cues from the previous subject, hindering complete subject replacement. A naive approach to addressing this issue is to derive a subject mask from cross-attention maps (Cao et al., 2023) and use it to selectively update the cache. However, subject-related attention may also respond to surrounding background regions, making it difficult to precisely isolate the cache regions that require updating. As illustrated in Figure 5, subtracting appropriately weighted background attention suppresses background responses while preserving subject-related responses. This observation suggests using subject–background attention contrast to obtain cleaner spatial guidance for selective cache updates.

Building on this observation, we introduce SubjectSwitch, a training-free framework that selectively updates the recent KV cache and controls the influence of the persistent attention sink. Figure 1 illustrates subject switching while maintaining background continuity across prompt updates. Given two consecutive prompts, SubjectSwitch first identifies subject changes, then localizes subject-related cache regions by subtracting appropriately weighted background attention from subject attention. This suppresses background responses and produces masks that distinguish regions requiring reconstruction from those to be preserved. For subject replacement, we suppress the outgoing subject in the historical latents before recomputing the corresponding regional KV representations under the new prompt, while preserving the cached background representations. This reduces the risk of retaining the previous subject's appearance in the reconstructed cache despite the updated text condition. The key distinction from feature-based selective forgetting is the use of prompt-guided subject–background localization together with suppression of outgoing-subject cues in the historical latents used for cache reconstruction.

Cache reconstruction alone may not eliminate visual cues from the initial frames, as the persistent attention sink can continue to influence subsequent generation. To estimate its influence, we compare predictions with and without sink access at a subset of generation blocks and denoising timesteps, keeping the noisy latents, prompt, and all other conditions identical. The prediction differences indicate where and how strongly the sink affects the current output. We then assess whether this influence shows signs of reintroducing content from the initial frames and selectively correct predictions in the affected regions. To avoid abrupt transitions during prompt adaptation, we additionally blend predictions conditioned on the previous and current prompts, using separate schedules for foreground and background regions. Together, these mechanisms determine which cached regions to preserve or reconstruct, control the influence of the persistent attention sink, and regulate how the requested changes unfold over time.

On VidProM-Switch and MotionBind-Switch (400 sequences, 2,000 transitions), SubjectSwitch improves Clean Switch Rate over LongLive by 23.43 and 26.60 percentage points, respectively, while keeping all five VBench-Long scores within 1.3 points of LongLive. A user study shows 76%–94% preference over seven baselines (Appendix A).

Our main contributions are summarized as follows:

- We propose a training-free method for subject switching that uses subject–background attention contrast to localize subject-related cache regions and suppresses outgoing-subject cues in historical latents before selective KV cache reconstruction, while preserving cached background representations. We additionally control the influence of the persistent attention sink to reduce the reappearance of initial-frame content.
- We evaluate our method on VidProM-Switch and MotionBind-Switch, demonstrating improved subject adaptation across diverse interactive prompt transitions, complemented by a user study showing consistent human preference over seven baselines.

2 Related Work

Streaming and Autoregressive Video Generation. Diffusion Forcing unifies autoregressive and diffusion modeling, while SEINE, FIFO-Diffusion, and SkyReels-V2 extend video synthesis beyond fixed lengths (Chen et al., 2024a;b; Kim et al., 2024; Chen et al., 2025). CausVid and Self-Forcing improve causal streaming through distillation and autoregressive rollout training, respectively (Yin et al., 2025; Huang et al., 2025a). Rolling Forcing further reduces long-horizon error accumulation through joint denoising (Liu et al., 2026).

Interactive Generation and Memory Adaptation. LongLive combines KV-recache with a persistent frame sink (Yang et al., 2026), while Infinity-RoPE manages context through cache retention and RoPE adjustment (Yesiltepe et al., 2026). Deep Forcing introduces sink-based cache compression (Yi et al., 2025), and LongLive-RAG retrieves relevant historical latents (Hu et al., 2026). Grounded Forcing and Echo-Forcing adapt memory through temporal weighting and feature-based forgetting, respectively (Chen et al., 2026; Wu et al., 2026). Our focus is on selectively updating subject-related memory while preserving compatible background context.

Attention-Based Semantic Control. Prompt-to-Prompt and MasaCtrl enable image editing through cross-attention and mutual self-attention control, respectively (Hertz et al., 2023; Cao et al., 2023). Video-P2P and FateZero extend attention control to temporally coherent video editing (Liu et al., 2024; Qi et al., 2023). These methods primarily target offline editing, whereas we use subject–background attention contrast to guide selective reconstruction of historical KV caches during interactive prompt transitions.

3 Method

Subject-centric prompt transitions in interactive long-video generation require adapting to new subjects while retaining the background context that users intend to preserve. However, uniform cache re-encoding can retain obsolete subject cues alongside reusable background context. We present SubjectSwitch as a unified framework that selectively updates the recent KV cache and controls the influence of the persistent Frame Sink across such transitions (Figure 4). We first present the preliminaries and observations in Section 3.1, then introduce Subject-Aware KV Cache Routing in Section 3.2 and Counterfactual Sink Replay Suppression in Section 3.3.

3.1 Preliminaries and Observations

Preliminaries. Let z_t denote the t-th latent video block and p_t its text prompt. At denoising timestep τ, the frozen pretrained generator F_θ predicts a clean block from the noisy latent x_t^τ and historical KV cache $\mathcal{C}_{t-1}$:

$$\widehat{z}_t^\tau = F_\theta(x_t^\tau, \tau, p_t; \mathcal{C}_{t-1}). \tag{1}$$

After denoising, the generated block is re-encoded to update the cache. At each self-attention layer,

$$\mathcal{C}_{t-1} = \mathcal{S} \oplus \mathcal{H}_{t-1}, \tag{2}$$

where $\mathcal{S}$ is a persistent Frame Sink retaining the initial block, $\mathcal{H}_{t-1}$ stores recent visual context, and $\oplus$ concatenates keys and values along the token dimension. Text encoding and layer indices are omitted for brevity.

When the prompt switches from p^- to p^+, the recent pre-switch latent blocks Z^{r} are re-encoded under the new prompt:

$$\mathcal{H}^+ = \mathcal{R}_\theta(Z^{\mathrm{r}}, p^+), \tag{3}$$

where $\mathcal{R}_\theta$ is the cache re-encoding operator induced by F_θ. The Frame Sink $\mathcal{S}$ remains available throughout subsequent generation.

Using LongLive (Yang et al., 2026) as our base generator, we analyze two memory-related failure modes during subject transitions to motivate our method.



A little boy in a blue T-shirt transforms into Iron Man.

Figure 2: Subject mixing under uniform cache re-encoding in LongLive-based generation, with overlapping figures and duplicated limbs.

Observation 1: Spatial and semantic limitations of recaching. Uniform cache re-encoding reconstructs the recent KV cache across the complete spatial grid without explicitly distinguishing the edited subject from the surrounding background. Consequently, adapting to the new prompt can also modify background representations outside the intended edit region. Spatially selective updates address this mismatch, but do not by themselves ensure complete removal of source-subject information: the recomputed KV representations still derive from historical latents depicting the source subject, even when conditioned on the target prompt. Thus, fully replacing selected subject-associated entries can retain source appearance cues within the updated region.

Figure 2 shows incomplete subject replacement, with residual source appearance and mixed subject features. These limitations motivate controlling both the update region and the visual content used for cache reconstruction.

We next examine the second memory pathway, namely the persistent Frame Sink, and visualize its behavior after a subject transition.

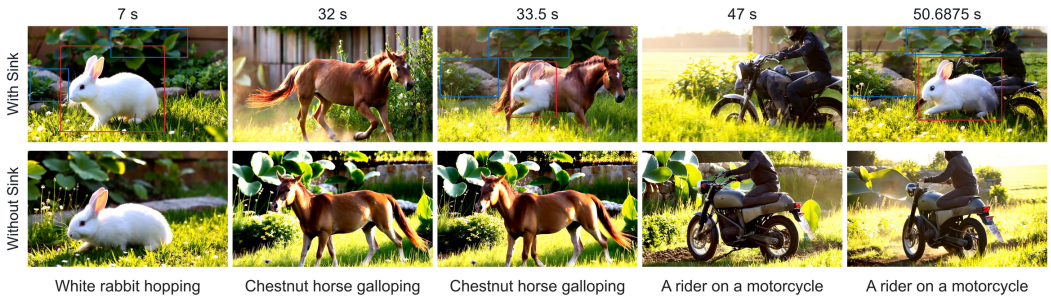


Figure 3: Initial-frame recurrence in LongLive-based generation. Top: Frame Sink enabled. Bottom: Sink access disabled from the start. Both runs use identical prompts, seeds, and sampling settings with a nine-frame rolling cache. Red and blue boxes mark recurrence of the initial subject and background appearance, respectively.

Observation 2: Persistent memory can become semantically outdated. The Frame Sink provides a stable visual reference, but retains initial-subject information after the corresponding frames

leave the rolling window. Useful background references and potentially conflicting subject information therefore coexist within the same memory. Such cues may contribute to initial-subject recurrence, as illustrated in Figure 3. Removing the entire Sink also discards useful background context, motivating selective control of its influence.

Together, these observations motivate selective subject updates with source-subject suppression during cache reconstruction, and selective Sink control that retains useful background context. SubjectSwitch addresses these requirements through subject-aware KV cache routing and counterfactual Sink replay suppression.

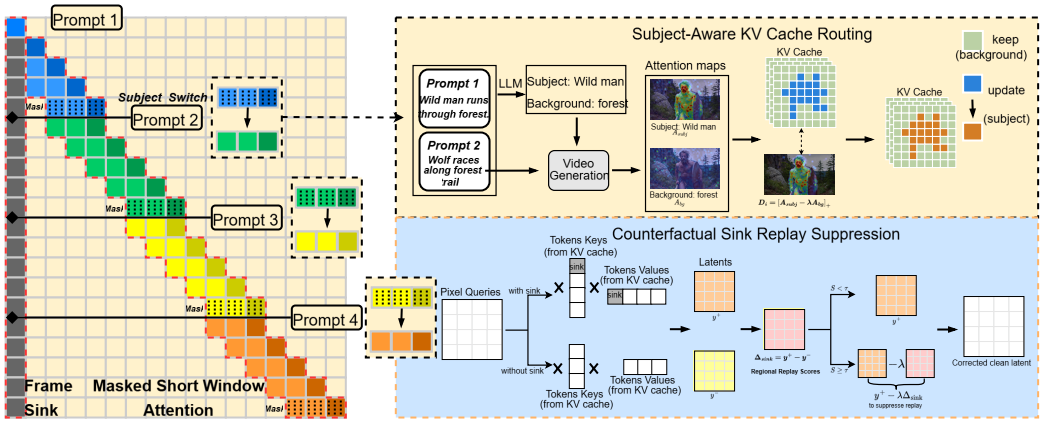


Figure 4: **Overview of SubjectSwitch.** Left: Frame Sink and masked short-window attention support generation across prompt transitions. Top right: Subject-aware KV cache routing uses subject–background attention contrast to derive masks and selectively update subject-related KV entries while preserving background entries. Bottom right: Counterfactual Sink replay suppression compares predictions with and without Sink access to guide selective prediction correction. Our method operates only on a subset of blocks and timesteps, thus incurring negligible additional computational overhead.

3.2 Subject-Aware KV Cache Routing



Figure 5: Background suppression for subject localization. (a) Generated frame. (b) Subject attention. (c) Background attention. (d) Positive attention difference. (e–f) Diagnostic masks from (b) and (d) using the same threshold. Attention maps share a common color scale.

Subject-centric prompt transitions change the subject while retaining the background, e.g., replacing a "wild man" with a "wolf" in a forest. Unlike uniform cache re-encoding, which updates both regions indiscriminately, we localize the source subject and selectively replace its cached keys and values, preserving background entries.

Offline LLM-based parsing with DeepSeek-V4-Flash (DeepSeek-AI et al., 2026) provides precomputed semantic differences between adjacent prompts, identifying edited subjects and stable background terms. We probe pre-switch latents under the previous prompt p^-, aggregating subject and background cross-attention over the same selected layers. Attribute and part edits use their shared host for localization. Let A_i^{subj} and A_i^{bg} denote these maps for frame i. To suppress background interference before mask extraction (Fig. 5), we compute

$$D_i = \left[A_i^{\text{subj}} - \lambda A_i^{\text{bg}}\right]_+, \qquad M_i^{(0)} = \mathcal{P}_\rho(\text{Norm}(D_i)) \,. \tag{4}$$

Here, $[\cdot]_+$ retains positive responses, λ controls background suppression, and Norm performs per-frame spatial min–max normalization. $\mathcal{P}_\rho$ selects high-response tokens toward cumulative attention-

mass fraction ρ, subject to minimum and maximum area bounds. If selection fails quality checks, we attempt bounded recovery from the same attention evidence. Cycle-consistent correspondences from cached self-attention value features refine these masks across frames; binarization and area clipping yield the routing masks $M_i \in \{0,1\}^{H' \times W'}$ on the spatial token grid.

Before re-encoding, we replace masked source regions in the pre-switch recaching input with their per-frame, per-channel background mean, leaving saved video latents and target frames unchanged. Re-encoding this context $\widetilde{Z}^{\mathrm{r}}$ under the new prompt p^+ produces candidate cache $\mathcal{H}^{\mathrm{rec}} = \mathcal{R}_\theta(\widetilde{Z}^{\mathrm{r}}, p^+)$. At each layer, we update

$$\widetilde{K}_i = (1 - M_i) \odot K_i^- + M_i \odot K_i^{\mathrm{rec}}, \qquad \widetilde{V}_i = (1 - M_i) \odot V_i^- + M_i \odot V_i^{\mathrm{rec}}. \tag{5}$$

Superscripts $-$ and rec denote original and candidate entries; $\odot$ denotes element-wise multiplication, broadcasting masks across heads and channels. Subject entries are fully replaced; background entries are retained.

Visible transitions are controlled separately: old- and new-prompt branches receive identical noisy latents and denoising timesteps, and read the same routed recent KV memory, without committing branch-specific updates. Their clean-latent predictions are blended using attention-derived spatial support and soft weights. In our configuration, monotonic subject and background blending schedules reach one by the end of the first and second blocks, respectively.

After the first target block, available target terms localize its subject under p^+. We merge source and target visual supports by union and their soft weights by element-wise maximum. A second recaching pass fully commits target frames while updating pre-switch history only through the original source masks M_i. The merged visual support therefore does not replace the source masks governing historical cache writes. Pure subject additions preserve history, skip initial source recaching and blending, and defer localization until the first target block, whose frames are then committed in full.

Although this updates recent memory, obsolete subject cues can persist in the Frame Sink, motivating the counterfactual replay suppression introduced next.

3.3 Counterfactual Sink Replay Suppression

Following the lower branch of Figure 4, we compare predictions with and without Sink access, assess replay, and conditionally correct subject and background regions.

Let $\mathcal{S}^0$ and Z^0 denote the fixed initial clean Sink cache and latent block. At selected block–timestep pairs (t, τ), two branches share noisy input x_t^τ, prompt p_t, and recent cache $\mathcal{H}_{t-1}$:

$$y_t^+ = F_\theta\big(x_t^\tau, \tau, p_t; \mathcal{S}^0 \oplus \mathcal{H}_{t-1}\big), \quad y_t^- = F_\theta(x_t^\tau, \tau, p_t; \mathcal{H}_{t-1}), \quad \Delta_t = y_t^+ - y_t^-. \tag{6}$$

Here, $+$ and $-$ denote full-Sink and no-Sink clean-latent predictions, with dependence on τ omitted. Both branches leave historical KV entries unchanged and differ only in access to the initial Sink keys and values. The difference Δ_t measures Sink influence, which may include useful background context as well as replay.

Replay is evaluated on source-subject support M_t^{src} inherited from routing and background $B_t = 1 - M_t^{\mathrm{qry}}$. The query mask initially uses source support and switches to target support once localized. The source support remains available for subject correction, while the updated query mask protects the localized target from background correction. Each regional score r_t^u compares appearance and temporal-difference motion similarities to Z^0 and the preceding block z_{t-1}, and incorporates boundary discontinuity and positive cosine alignment between Δ_t and $Z^0 - z_{t-1}$. For $u \in \{\mathrm{sub}, \mathrm{bg}\}$, we define

$$g_t^u = \mathbf{1}[r_t^u \geq \eta_u] \,\mathrm{sigmoid}\left(\frac{r_t^u - \eta_u}{T}\right), \tag{7}$$

with regional threshold η_u and temperature $T > 0$. The two gates are evaluated independently, allowing subject recurrence to be suppressed even when background continuity is preserved. The corrected clean-latent prediction is

$$\widehat{y}_t = y_t^+ - \gamma_{\mathrm{sub}} g_t^{\mathrm{sub}} M_t^{\mathrm{src}} \odot \Delta_t - \gamma_{\mathrm{bg}} g_t^{\mathrm{bg}} \Pi^+_{d_t^{\mathrm{bg}}}(B_t \odot \Delta_t), \tag{8}$$

where γ_{sub} and γ_{bg} control correction strengths, $d_t^{\text{bg}} = B_t \odot (Z^0 - z_{t-1})$, and

$$\Pi_d^+(v) = \text{clip}\left(\frac{\langle v, d\rangle}{\max(\|d\|_2^2, \epsilon)}, 0, c_{\max}\right) d.$$

The subject term removes the gated Sink contribution within its support, whereas background correction acts only along the reference replay direction. The coefficient is capped by $c_{\max}$, and $\epsilon > 0$ ensures numerical stability. When both gates are zero, $\widehat{y}_t = y_t^+$.

After denoising, the accepted block is re-encoded as clean context. The latest gates select full-Sink access when neither activates, background-only Sink access for subject-only replay, and no-Sink access when background replay activates. Token exclusion uses an initial-frame mask, distinct from prediction masks; $\mathcal{S}^0$ remains unchanged.

Our configuration evaluates the first two denoising steps of alternating blocks, starting with the first active block, requiring one additional no-Sink pass per selected pair regardless of gate activation. Skipped pairs use full-Sink predictions without correction; only gates carry over to clean-context re-encoding. The schedule and gates reset at prompt switches or reactivation.

During the visual transition in Section 3.2, correction applies to the new-prompt branch before spatial blending, with $\gamma_{\text{bg}} = 0$ and background-only Sink access during clean-context re-encoding. Target-only additions bypass this mechanism.

4 Experiment

4.1 Experimental Setups

Implementation. We build SubjectSwitch upon LongLive (Yang et al., 2026) without additional training or fine-tuning. We generate approximately 60-second videos following six sequential prompts.

Datasets and evaluation Metrics. We construct **VidProM-Switch** and **MotionBind-Switch** from VidProM (Wang & Yang, 2024) and the motion-binding prompts of T2V-CompBench (Sun et al., 2025), respectively. Each dataset contains 200 sequences of six prompts. We report five VBench-Long (Huang et al., 2024; 2025b) dimensions (SC, BC, MS, AQ, and IQ), three subject-switching metrics (SSC, CSR, and SLS), and mean reversion severity (RS). Higher scores are better for all metrics except RS. Implementation details, dataset construction procedures, and evaluation protocols are provided in Appendix B.

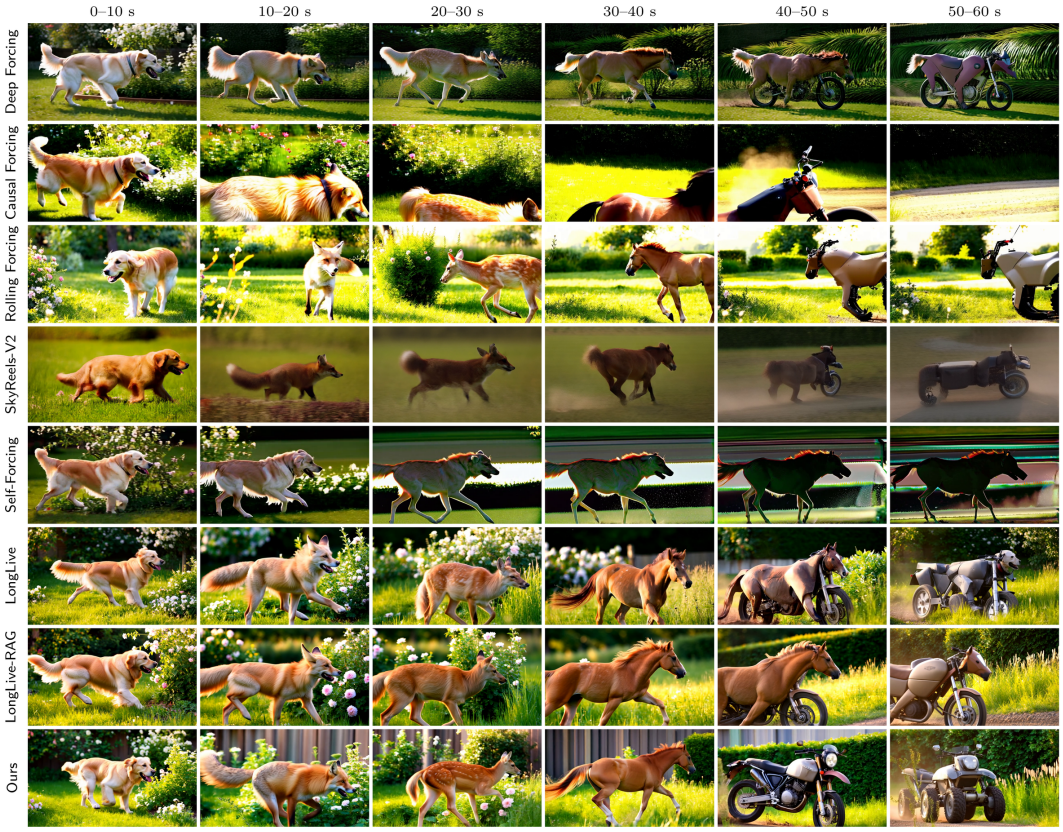
4.2 Comparison with Other Methods

Qualitative comparison. Figure 6 compares generation across six sequential prompts. Several baselines show incomplete subject replacement: Self-Forcing retains animal-like subjects after vehicle prompts, while Deep Forcing, LongLive, and LongLive-RAG combine animal structures with vehicle components. SubjectSwitch more clearly follows the requested sequence, producing distinct motorcycle and rover structures while retaining compatible background context. Additional comparisons are provided in Appendix C.1.

Quantitative results. Table 1 shows that SubjectSwitch achieves the highest SSC, CSR, and SLS on both datasets, indicating cleaner subject replacement and earlier clean target-subject emergence. CSR improves over LongLive (Yang et al., 2026) by 23.43 and 26.60 percentage points on VidProM-Switch and MotionBind-Switch, respectively. These gains accompany decreases of at most 1.3 points in the five reported VBench-Long scores. RS also decreases substantially relative to LongLive, although several baselines achieve lower RS with weaker subject-switching performance. Metric definitions and sampling settings are provided in Appendix B.3.

Table 1: Quantitative comparison on VidProM-Switch (V) and MotionBind-Switch (M). Each cell reports V / M scores on a 0–100 scale. Bold and underlined values indicate the best and second-best results within each dataset, respectively.

	Switch metrics (V / M)				VBench-Long metrics (V / M)				
Method	SSC ↑	CSR ↑	SLS ↑	RS ↓	SC ↑	BC ↑	MS ↑	AQ ↑	IQ ↑
LongLive	70.43 / 76.72	54.92 / 61.06	56.50 / 53.67	27.50 / 33.47	95.10 / 95.96	94.76 / 95.29	98.49 / 98.81	62.14 / 56.14	68.32 / 69.25
Deep Forcing	68.70 / 73.55	49.02 / 51.60	45.90 / 42.71	4.28 / 4.39	95.19 / 95.96	94.75 / 95.24	98.21 / 98.50	61.21 / 55.03	68.51 / 67.78
Rolling Forcing	51.90 / 66.10	22.44 / 36.81	18.47 / 24.01	0.45 / 0.18	96.25 / 97.17	95.38 / 96.04	98.53 / 98.75	61.47 / 54.99	**71.13 / 70.25**
Causal-Forcing	58.44 / 77.16	23.82 / 39.68	13.71 / 23.26	0.42 / 0.15	94.24 / 94.22	93.69 / 94.27	97.21 / 97.72	58.62 / 51.45	68.97 / 65.98
Self-Forcing	47.28 / 63.95	26.18 / 31.17	25.23 / 28.10	**0.13** / 1.61	95.87 / **97.49**	95.55 / 96.36	98.27 / 99.02	56.87 / 51.71	66.41 / 65.73
LongLive-RAG	65.90 / 75.71	47.83 / 56.28	52.40 / 46.93	0.16 / 0.24	**96.28** / 97.16	95.64 / 96.06	98.31 / 98.61	**62.63 / 57.38**	68.57 / 69.04
SkyReels-V2	63.36 / 79.92	27.95 / 43.40	17.68 / 30.78	1.39 / **0.07**	94.38 / 96.66	**95.73 / 96.67**	**98.94 / 99.20**	56.36 / 51.13	51.73 / 51.53
SubjectSwitch (Ours)	**85.52 / 92.70**	**78.35 / 87.66**	**74.08 / 76.17**	0.45 / 6.24	94.98 / 95.26	94.73 / 94.90	98.29 / 98.63	60.88 / 54.92	67.71 / 68.43



0–10 s	A golden retriever scampers rightwards across a lush garden in warm late-afternoon sunlight.
10–20 s	A red fox darts rightwards through the garden's grassy edge, its brush streaming in warm sunlight.
20–30 s	A fallow deer bounds rightwards across a broad meadow through knee-high grass in warm sunlight.
30–40 s	A chestnut horse gallops rightwards across the open pasture, mane streaming in golden sunlight.
40–50 s	A scrambler motorcycle speeds rightwards along a dirt track, kicking up dust in warm sunlight.
50–60 s	A six-wheeled rover travels rightwards along a gravel road toward a research facility in warm light.

Figure 6: Qualitative comparison of 60-second interactive video generation. Rows correspond to methods and columns to successive prompt intervals. Prompt summaries are shown below.

4.3 Ablation Study

We incrementally add subject-aware KV cache routing (SAR) and counterfactual Sink replay suppression (CSRS) to LongLive to obtain the full SubjectSwitch method (Table 2). SAR provides the main gains in SSC, CSR, and SLS in this incremental study, but increases RS. Adding CSRS on top

of SAR reduces RS, with mixed effects on the three subject-switching metrics. Additional results and detailed analysis are provided in Appendix C.2.

Table 2: Incremental ablation of SubjectSwitch on VidProM-Switch (V) and MotionBind-Switch (M). Each cell reports V / M scores on a 0–100 scale. Bold and underlined values indicate the best and second-best results within each dataset, respectively.

	Switch and replay metrics (V / M)				VBench-Long metrics (V / M)				
Configuration	SSC ↑	CSR ↑	SLS ↑	RS ↓	SC ↑	BC ↑	MS ↑	AQ ↑	IQ ↑
LongLive	70.43 / 76.72	54.92 / 61.06	56.50 / 53.67	27.50 / 33.47	**95.10 / 95.96**	**94.76 / 95.29**	**98.49 / 98.81**	**62.14 / 56.14**	**68.32 / 69.25**
LongLive + SAR	**86.84** / 92.20	**78.94** / 86.81	74.02 / 74.17	34.06 / 53.36	94.63 / 94.80	94.63 / 94.85	98.47 / 98.76	62.02 / 55.90	68.16 / 68.83
LongLive + SAR + CSRS	85.52 / **92.70**	78.35 / **87.66**	**74.08 / 76.17**	**0.45 / 6.24**	94.98 / 95.26	94.73 / 94.90	98.29 / 98.63	60.88 / 54.92	67.71 / 68.43

4.4 User Study

We compare SubjectSwitch with seven baselines, collecting 50 responses per baseline comparison. Participants evaluate transition completeness, transition naturalness, visual quality, background stability, and instruction following. For each response, an overall preference for SubjectSwitch is recorded when it is preferred in at least three of the five dimensions.

Figure 7 reports overall preference rates and vote counts for each dimension. SubjectSwitch achieves overall preference rates of 76%–94% across the seven comparisons, including 78% against LongLive. The study interface is provided in Appendix A (Figure 8).

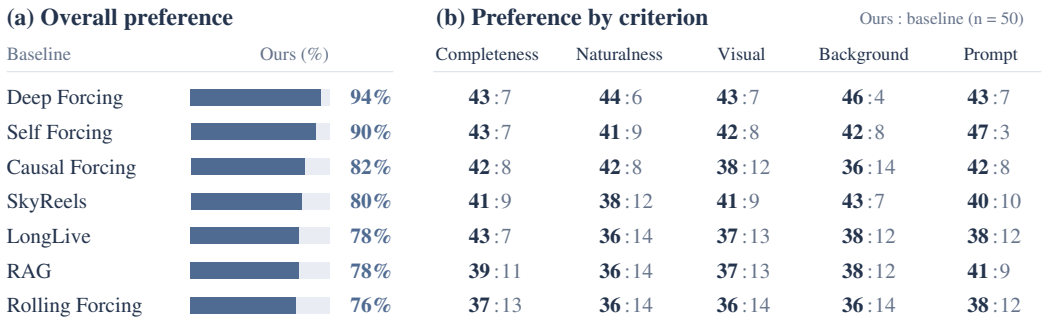


Figure 7: Human preference over seven baselines.

5 Conclusion

We presented SubjectSwitch, a training-free framework for subject switching in interactive long-video generation that selectively updates the recent KV cache and controls the influence of the persistent Frame Sink. Background suppression in subject attention localizes the source subject for subject-aware KV cache routing. Source-subject cues are suppressed in reconstruction inputs, and subject-related cache entries are selectively replaced while compatible background entries are retained. Counterfactual comparisons of predictions with and without Frame Sink access guide regional prediction correction and Sink access during context re-encoding. On VidProM-Switch and MotionBind-Switch, SubjectSwitch achieves the highest SSC, CSR, and SLS scores among the compared methods, improving Clean Switch Rate over LongLive by 23.43 and 26.60 percentage points, respectively, while keeping all five reported VBench-Long scores within 1.3 points of LongLive. These results support selective adaptation of visual history as an effective approach to balancing prompt responsiveness and scene continuity.

AI USE STATEMENT

We used generative AI to assist with language polishing and literature discovery, and to construct evaluation prompts. Literature discovery focused on streaming and autoregressive video generation, interactive generation and memory adaptation, and attention-based semantic control, including representative work such as Diffusion Forcing, LongLive, Infinity-RoPE, Prompt-to-Prompt, and Video-P2P. The authors reviewed the sources and finalized the discussion. For the subject-switching evaluation, DeepSeek-V4-Flash expanded each source prompt into six consecutive prompts for each of the two datasets, yielding 200 sequences per dataset and 400 in total, with 2,000 candidate transitions overall. The reported metrics include only transitions identified as changes in the main subject's identity or category. Details are described in the paper.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22560–22570, 2023.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024a.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. URL https://arxiv.org/abs/2504.13074.

Jintao Chen, Chengyu Bai, Junjun Hu, Xinda Xue, and Mu Xu. Grounded Forcing: Bridging time-independent semantics and proximal dynamics in autoregressive video synthesis. *arXiv preprint arXiv:2604.06939*, 2026. URL https://arxiv.org/abs/2604.06939.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: Short-to-long video diffusion model for generative transition and prediction. In *International Conference on Learning Representations*, 2024b.

DeepSeek-AI et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

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. URL https://arxiv.org/abs/2606.02553.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 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*, pp. 21807–21818, 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.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. FIFO-Diffusion: Generating infinite videos from text without training. In *Advances in Neural Information Processing Systems*, volume 37, pp. 89834–89868, 2024.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IAyzXjbfwo.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-P2P: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8599–8608, 2024.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025.

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

Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65618–65642, 2024. URL https://openreview.net/forum?id=pYNl76onJL.

Mingqiang Wu, Weilun Feng, Zhefeng Zhang, Haotong Qin, Yuqi Li, Guoxin Fan, Xiaokun Liu, Zhulin An, Libo Huang, Yongjun Xu, and Chuanguang Yang. Echo-Forcing: A scene memory framework for interactive long video generation. *arXiv preprint arXiv:2605.16003*, 2026. URL https://arxiv.org/abs/2605.16003.

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. In *International Conference on Learning Representations*, 2026.

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. URL https://arxiv.org/abs/2512.05081.

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

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. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02214.

A User Study

We compare SubjectSwitch with seven baselines, collecting 50 responses per baseline comparison. Participants evaluate transition completeness, transition naturalness, visual quality, background stability, and instruction following. Each response records a preference for one of the two videos in each dimension. For each response, an overall preference for SubjectSwitch is recorded when it is preferred in at least three of the five dimensions. Figure 8 shows the study interface.

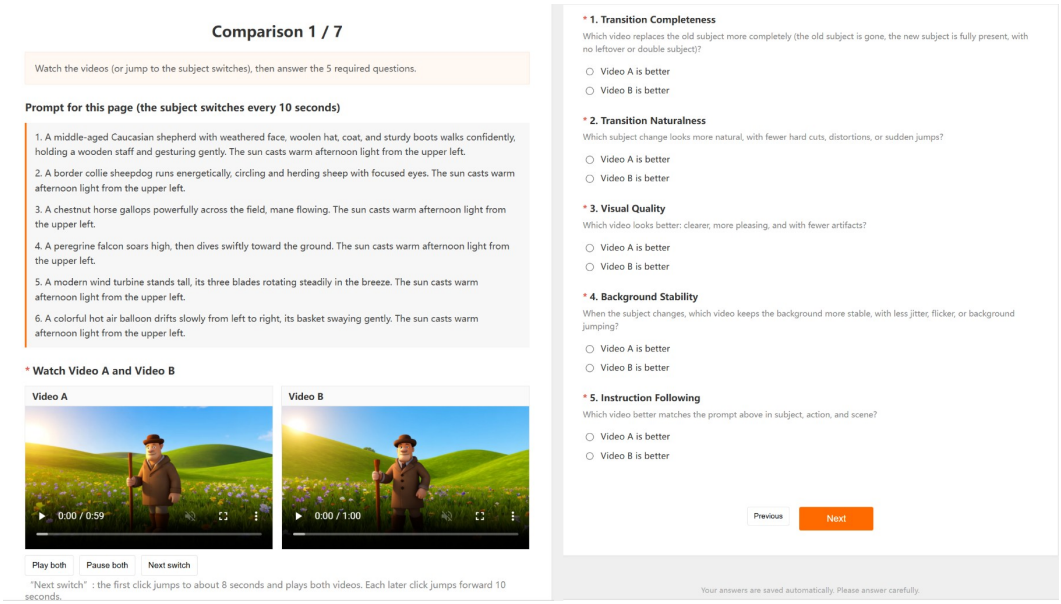


Figure 8: User study interface.

B Detailed Experimental Setups

B.1 Implementation Details

Backbone and hardware. We build SubjectSwitch upon LongLive (Yang et al., 2026), which is based on Wan2.1-T2V-1.3B (Team Wan et al., 2025). We use its pretrained generator and LoRA adapter without additional training or fine-tuning. Experiments are conducted on a single NVIDIA A40 GPU using `bfloat16` precision and a batch size of one.

Generation and cache configuration. Videos are generated at 832×480 resolution and 16 FPS, with four denoising steps per three-latent-frame block. Each video contains 240 latent frames, corresponding to approximately 60 seconds, and follows six sequential prompts with switches aligned to block boundaries. The KV-cache budget comprises three initial frames in the Frame Sink and nine rolling frames, with relative-window rotary position embeddings enabled.

Subject-aware cache adaptation. Subject differences are precomputed using `deepseek-v4-flash` and cached offline with a minimum confidence threshold of 0.85. We aggregate cross-attention maps from eight selected layers to localize subject regions and selectively update their KV entries while preserving cached background entries. During transitions, the source- and target-prompt branches share the same noisy latents. Subject blending follows $[0.15, 0.70, 1.00]$ over one block, while background blending extends over two blocks. We additionally enable counterfactual Sink replay suppression through paired predictions with and without Sink access.

B.2 Dataset Construction

We construct two prompt datasets, **VidProM-Switch** and **MotionBind-Switch**, to evaluate subject transitions in interactive long-video generation.

VidProM-Switch. We select 200 prompts from VidProM (Wang & Yang, 2024), each containing clearly specified subjects. The resulting dataset covers varied subject and background descriptions, providing a diverse setting for evaluating subject transitions.

MotionBind-Switch. We use all prompts in `prompts/4_motion_binding.txt`, the motion-binding prompt file of T2V-CompBench (Sun et al., 2025). This dataset focuses on subject transitions under prompts derived from explicit motion directions and relationships between subjects and their movements.

Prompt-sequence construction. For both datasets, we use `deepseek-v4-flash` (DeepSeek-AI et al., 2026) to expand each source prompt into six consecutive prompts. This produces 200 prompt sequences per dataset, each containing five candidate prompt transitions. The two datasets therefore contain 400 sequences and 2,000 candidate transitions in total. The subject-switching metrics below include only transitions identified as changes in the main subject's identity or category.

B.3 Evaluation Metrics and Protocol

General video quality. We report five dimensions from VBench-Long (Huang et al., 2024; 2025b): subject consistency (SC), background consistency (BC), motion smoothness (MS), aesthetic quality (AQ), and imaging quality (IQ). We additionally report Subject Switch Cleanliness (SSC), Clean Switch Rate (CSR), Switch Latency Score (SLS), and mean reversion severity (RS). The parameter values below specify the default configuration of the two evaluation scripts.

Subject-transition assessment. We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) to identify the source and target subjects from consecutive prompts and determine whether the main subject's identity or category changes. Changes limited to action, camera, lighting, style, background, or minor attributes are excluded unless they also change subject identity or category. For transformation prompts, the final requested subject is treated as the target. Subject-pair extraction and frame classification use `do_sample=False`, with at most 256 new tokens per response.

For video v, let $T_v = N_v/f_v$ be its decoded duration, where N_v is the frame count and f_v is the average frame rate. Unless an explicit segment duration is supplied, the evaluator uses nominal switch times $s_{v,k} = kT_v/6$ for $k = 1, \ldots, 5$. For each eligible transition e, frames are sampled at offsets $0.5, 1, 1.5, 2, 3, 5, 7, 9$ seconds after its nominal switch time. Offsets reaching the next prompt interval or the video end are excluded. A source-subject reference frame is taken 0.5 seconds before the switch, with its time clipped at zero. Requested times are mapped to the nearest decoded frame. Each classification receives the current frame, the reference frame, both prompts, and the extracted source and target subjects.

The judge assigns one of four labels: `clean_new`, `old_leakage`, `hybrid`, or `new_missing_or_unclear`, and reports whether the source subject, target subject, and hybrid traits are visible. Let $c_{e,j} \in \{0, 1\}$ denote contamination at sample j: it is one if the label is `old_leakage` or `hybrid`, or if source-subject visibility or hybrid traits are reported. Let $q_{e,j} \in \{0, 1\}$ denote clean target presence: it is one only when the label is `clean_new`, the target subject is visible, and neither the source subject nor hybrid traits are reported. Missing or unclear target frames remain in the evaluation; they do not count as clean target frames and count as contamination only if source-subject or hybrid evidence is present. The judge's confidence is recorded but does not filter or weight the judgments.

Subject Switch Cleanliness (SSC). Let $\mathcal{E}$ be the set of eligible subject-changing transitions in a dataset, and let m_e be the number of valid sampled frames for transition e. SSC combines the occurrence and sampled frequency of contamination:

$$\mathrm{SSC} = \frac{100}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[1 - \omega \max_j c_{e,j} - (1 - \omega) \frac{1}{m_e} \sum_{j=1}^{m_e} c_{e,j} \right], \tag{9}$$

where $\omega = 0.6$. SSC measures the absence of source-subject contamination; clean target emergence is additionally assessed by CSR and SLS.

Clean Switch Rate (CSR). CSR requires no sampled contamination, timely clean target emergence, and sustained clean target presence. Let $d_{e,j}$ be the offset of sample j after the switch. The clean-emergence latency is

$$L_e = \min\{d_{e,j} : j < m_e,\ d_{e,j} \le L_{\max}, \quad q_{e,j} = q_{e,j+1} = 1\}, \tag{10}$$

where $L_{\max} = 3$ seconds and the minimum of an empty set is $+\infty$. Thus, two consecutive sampled frames must show a clean target. The three-second limit applies to the first sample of the pair; the confirming sample may occur later.

The stable window contains samples at offsets of at least two seconds. Let $\mathcal{W}_e = \{j : d_{e,j} \ge 2\}$ and $R_e = |\mathcal{W}_e|^{-1} \sum_{j \in \mathcal{W}_e} q_{e,j}$. If this window is empty, all valid samples are used instead. The dataset-level score is

$$\mathrm{CSR} = \frac{100}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbf{1}\left[\max_j c_{e,j} = 0 \ \wedge\ L_e < \infty \ \wedge\ R_e \ge 0.75 \right]. \tag{11}$$

Switch Latency Score (SLS). SLS rewards earlier clean target emergence using the same latency L_e:

$$\mathrm{SLS} = \frac{100}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[1 - \frac{\min(L_e, L_{\max})}{L_{\max}} \right]. \tag{12}$$

A transition without a qualifying clean pair receives zero. Because the earliest sampled offset is 0.5 seconds, the maximum attainable SLS under this sampling schedule is $100(1 - 0.5/3) \approx 83.33$. No additional rescaling is applied.

Mean reversion severity (RS). RS measures cumulative severity of suspected abrupt returns toward the first frame's appearance. It is computed independently of prompt annotations and Qwen judgments, using all decoded frames at their native resolution and frame rate. The default configuration uses all three color channels without grayscale conversion or initial-frame exclusion. Let $\mathbf{x}_{v,n}$

denote the flattened pixel vector of frame n in video v. We compute cosine similarity to the first decoded frame and its change between adjacent frames:

$$a_{v,n} = \frac{\langle \mathbf{x}_{v,n}, \mathbf{x}_{v,1} \rangle}{\|\mathbf{x}_{v,n}\|_2 \|\mathbf{x}_{v,1}\|_2}, \qquad \delta a_{v,n} = a_{v,n} - a_{v,n-1}. \tag{13}$$

If both pixel vectors have zero norm, their similarity is one; if only one has zero norm, it is zero. A suspected return is detected only when $\delta a_{v,n} > \theta_\Delta$ and $a_{v,n} > \theta_A$, where $\theta_\Delta = 0.02$ and $\theta_A = 0.85$. Its contribution is

$$\kappa_{v,n} = [\delta a_{v,n} - \theta_\Delta]_+ [a_{v,n} - \theta_A]_+, \tag{14}$$

where $[\cdot]_+ = \max(\cdot, 0)$. The cumulative penalty and per-video severity are

$$P_v = \sum_{n=2}^{N_v} \kappa_{v,n}, \qquad \mathrm{RS}_v = 100\,(1 - \exp(-\alpha P_v))\,, \tag{15}$$

with $\alpha = 100$. Each qualifying adjacent-frame transition contributes once; consecutive detections are not merged into episodes. The cumulative penalty is not normalized by video length or by the number of evaluated frame pairs. PSNR and adjacent-frame cosine similarity are recorded as diagnostics but do not enter RS.

Aggregation and interpretation. SSC, CSR, and SLS are pooled with equal weight over all eligible subject-changing transitions in each dataset, rather than by averaging per-video means. Videos without eligible subject changes do not contribute to these three scores. Their raw evaluator outputs lie in $[0, 1]$ and are multiplied by 100 for reporting. If no eligible transitions exist, these scores are undefined.

RS is the equal-weight mean of per-video severities:

$$\mathrm{RS} = \frac{1}{|\mathcal{V}_{\mathrm{ok}}|} \sum_{v \in \mathcal{V}_{\mathrm{ok}}} \mathrm{RS}_v, \tag{16}$$

where $\mathcal{V}_{\mathrm{ok}}$ contains videos with complete decoding and at least one valid adjacent-frame comparison. Videos reported as erroneous, incompletely decoded, or insufficient are excluded from this mean. RS is a pixel-based heuristic expressed on a 0–100 scale; it is not a probability of semantic subject recurrence. Higher scores are better for all reported metrics except RS.

C Additional Qualitative and Ablation Results

C.1 Detailed Comparison Results

This subsection supplements the main comparisons with additional qualitative examples and quantitative analysis. The supplementary visualizations correspond to the **upper panels (a)** of Figures 9, 10, 11, and 12. They complement Figure 6 and Table 1 by illustrating responses to sequential subject-changing instructions.

Comparison setting and visualization. We compare SubjectSwitch with seven baselines: LongLive, Deep Forcing, Rolling Forcing, Causal-Forcing, Self-Forcing, LongLive-RAG, and SkyReels-V2. Each example follows six sequential prompts over an approximately 60-second video. Rows represent methods, while columns show representative frames from successive 10-second intervals. The instructions are listed between panels (a) and (b). This layout supports comparisons across methods within each interval and inspection of subject evolution across prompt updates.

Subject replacement and instruction following. The qualitative comparison examines whether the target subject emerges without retaining incompatible characteristics of the previous subject. In the main-text example in Figure 6, Self-Forcing retains animal-like subjects after vehicle prompts, while Deep Forcing, LongLive, and LongLive-RAG combine animal structures with vehicle components. SubjectSwitch follows the requested sequence more clearly, producing distinct motorcycle and rover structures while retaining compatible background context. The supplementary upper panels provide further examples for inspecting target-subject appearance, residual source characteristics, and agreement with the instructions.

Subject-switching performance. Table 1 shows that SubjectSwitch achieves the highest SSC, CSR, and SLS among the compared methods on both datasets. Its SSC scores are 85.52 and 92.70 on VidProM-Switch and MotionBind-Switch, respectively. CSR reaches 78.35 and 87.66, improving over LongLive, the strongest baseline on this metric, by 23.43 and 26.60 percentage points. SLS reaches 74.08 and 76.17, exceeding LongLive by 17.58 and 22.50 points. These results indicate reduced source-subject contamination, more successful clean replacement, and earlier clean target emergence. Metric definitions and temporal sampling settings are provided in Appendix B.3.

Reversion severity. Relative to LongLive, SubjectSwitch reduces RS from 27.50 to 0.45 on VidProM-Switch and from 33.47 to 6.24 on MotionBind-Switch. Several baselines nevertheless obtain lower RS with weaker subject-switching performance. For example, LongLive-RAG achieves RS scores of 0.16 and 0.24, but CSR scores of only 47.83 and 56.28. RS should therefore be interpreted alongside the switching metrics, since low reversion severity alone does not establish successful subject replacement.

Background consistency and visual quality. The visual comparisons also examine whether subject changes preserve background content compatible with the current instruction. Relative to LongLive, SubjectSwitch obtains lower scores on the five reported VBench-Long dimensions, with maximum decreases of 1.26 points on VidProM-Switch and 1.22 points on MotionBind-Switch. The decreases in background consistency are 0.03 and 0.39 points, respectively. Thus, the switching improvements accompany measurable changes in general video quality, with all five score differences remaining below 1.3 points.

C.2 Detailed Ablation Results

This subsection analyzes the contributions of subject-aware KV cache routing (SAR) and counterfactual Sink replay suppression (CSRS). The qualitative results correspond to the **lower panels (b)** of Figures 9, 10, 11, and 12. Table 2 reports the corresponding quantitative results under the same approximately 60-second, six-prompt evaluation protocol.

Ablation configurations. We incrementally add SAR and CSRS to LongLive, yielding three configurations labeled "LongLive," "+SAR," and "Ours (Full)" in the lower panels. SAR selectively updates subject-related content in the recent KV cache, while CSRS further controls the influence

of the persistent Sink. The lower panels share the instruction sequence and temporal layout of the upper panels. Comparing the first two rows examines the contribution of SAR; comparing the last two examines the additional contribution of CSRS with SAR enabled.

Effect of SAR on subject switching. Adding SAR improves SSC, CSR, and SLS by 16.41, 24.02, and 17.52 points on VidProM-Switch, and by 15.48, 25.75, and 20.50 points on MotionBind-Switch. These improvements identify SAR as the main source of gains in clean subject replacement and switching responsiveness within the incremental study. They are consistent with its role in reducing obsolete subject information in the recent KV cache. However, RS increases from 27.50 to 34.06 on VidProM-Switch and from 33.47 to 53.36 on MotionBind-Switch. The switching gains from SAR alone therefore do not translate into reduced reversion severity.

Effect of CSRS on reversion suppression. Adding CSRS to the SAR-enabled configuration reduces RS from 34.06 to 0.45 on VidProM-Switch and from 53.36 to 6.24 on MotionBind-Switch. The full model achieves the lowest RS among the three ablation configurations on both datasets. These reductions are consistent with CSRS suppressing detected returns toward the initial frame's appearance. The "+SAR" and "Ours (Full)" rows in the lower panels complement this analysis by showing generated content with and without the additional Sink correction, including its evolution across later intervals.

Effect of CSRS on subject-switching metrics. The effects of CSRS on SSC, CSR, and SLS vary by dataset. On MotionBind-Switch, adding CSRS increases these metrics by 0.50, 0.85, and 2.00 points, respectively, yielding final scores of 92.70, 87.66, and 76.17. On VidProM-Switch, SSC and CSR decrease by 1.32 and 0.59 points, while SLS increases slightly from 74.02 to 74.08. The full model retains substantial switching gains over LongLive, while the most consistent additional benefit of CSRS is reduced RS across both datasets.

Effects on general video quality. Relative to SAR alone, adding CSRS increases SC by 0.35 and 0.46 points and BC by 0.10 and 0.05 points on VidProM-Switch and MotionBind-Switch, respectively. Meanwhile, MS decreases by 0.18 and 0.13 points, AQ by 1.14 and 0.98 points, and IQ by 0.45 and 0.40 points. LongLive retains the highest scores on all five reported VBench-Long dimensions among the three configurations. These differences characterize the video-quality trade-offs accompanying improved subject switching and reversion suppression.

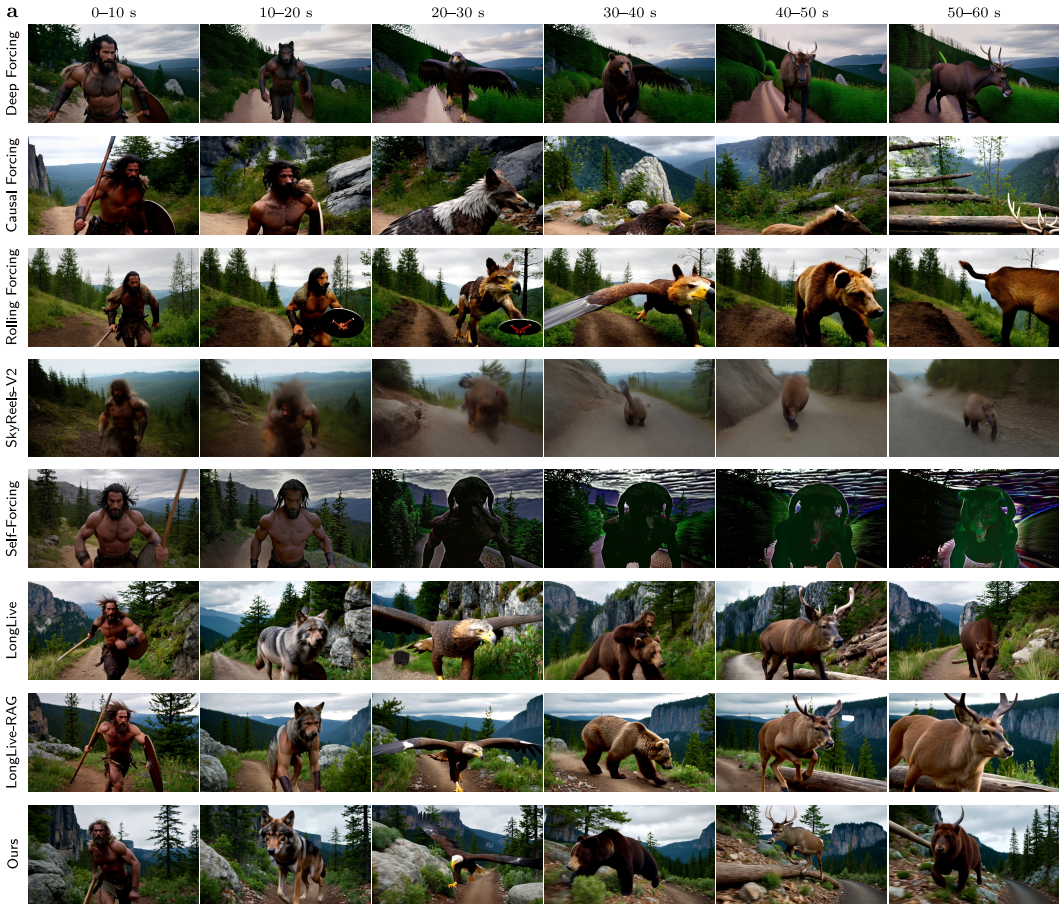


Figure 9: **Qualitative comparison and ablation results for Case 1.** **(a)** Comparison of SubjectSwitch with seven baseline methods. **(b)** Incremental ablation comparing LongLive, LongLive + SAR, and the full SubjectSwitch model (LongLive + SAR + CSRS). Columns follow the six successive 10-second intervals, with the corresponding instructions listed between the two panels.



Figure 10: **Qualitative comparison and ablation results for Case 2.** **(a)** Comparison of SubjectSwitch with seven baseline methods. **(b)** Incremental ablation comparing LongLive, LongLive + SAR, and the full SubjectSwitch model (LongLive + SAR + CSRS). Frames are arranged chronologically from left to right, and the text block specifies the instruction for each interval.

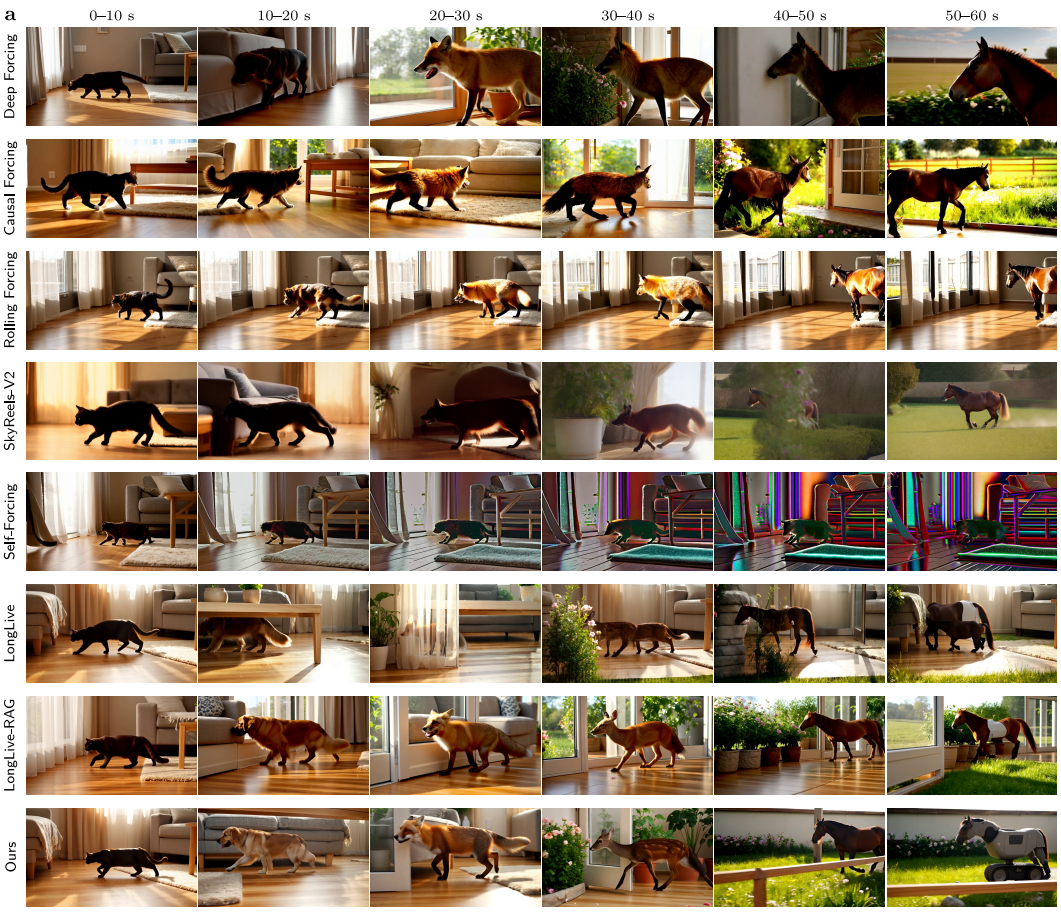


Figure 11: **Qualitative comparison and ablation results for Case 3.** **(a)** Comparison of SubjectSwitch with seven baseline methods. **(b)** Incremental ablation comparing LongLive, LongLive + SAR, and the full SubjectSwitch model (LongLive + SAR + CSRS). Each column shows the results for the same temporal interval across methods; the associated instructions are provided between the two panels.



Figure 12: **Qualitative comparison and ablation results for Case 4.** **(a)** Comparison of SubjectSwitch with seven baseline methods. **(b)** Incremental ablation comparing LongLive, LongLive + SAR, and the full SubjectSwitch model (LongLive + SAR + CSRS). The six columns summarize the video under successive instructions, allowing comparison of subject evolution and background consistency over time.