STREAMREAL: REFERENCE-GUIDED SUBJECT SWITCHING IN STREAMING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-guided subject switching requires a causal streaming generator to adopt subjects specified by new visual references during an ongoing rollout while preserving scene and temporal continuity. This is challenging because accumulated history entangles preceding-subject information with context needed for continuation. Comparisons on an adapted causal generator show that retaining the historical cache or recaching it under updated text yields weak alignment with new references, whereas clearing it improves responsiveness over subsequent chunks at the cost of aesthetics and motion smoothness. We introduce StreamReal, which organizes causal history to distinguish subject-dependent information from context that remains relevant across reference updates. To establish reference conditioning, Few-Step Conditioning Alignment aligns a pretrained reference pathway with a frozen few-step causal generator, followed by reference-conditioned Distribution Matching Distillation warm-up to adapt the reference branch to self-generated histories. During streaming, Memory Flow maintains a cross-chunk semantic index associating historical information with the evolving subject. Upon a reference update, Memory-Guided History Routing uses this index to regulate access to native historical states, suppressing preceding-subject evidence while preserving context for coherent continuation. We introduce a Subject Switch Benchmark spanning diverse subject categories and switching positions. Evaluations on OpenS2V-Eval and this benchmark show that StreamReal combines subject fidelity and visual quality with a favorable balance between switching responsiveness and temporal continuity.

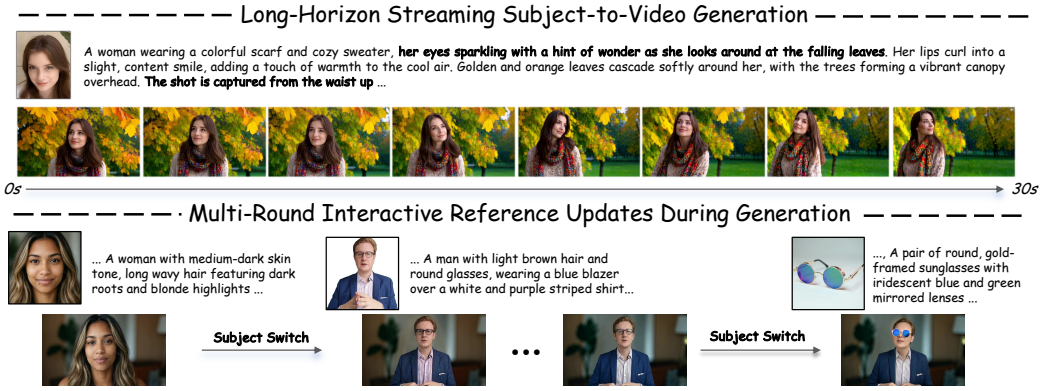


Figure 1: **StreamReal** achieves strong results on two challenging tasks: (1) long-horizon subject-to-video generation with high visual quality; and (2) multi-round reference updates during streaming generation while preserving scene context.

1 INTRODUCTION

Recent advances in causal video generation (Singer et al., 2022; Guo et al., 2023; Hong et al., 2022; Blattmann et al., 2023; Wu et al., 2023; Wang et al., 2023; Ho et al., 2022) are turning video synthesis from an offline process into an interactive streaming experience, where users can modify

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.

generation conditions while a video is being produced. Beyond changing textual instructions or motion controls, a natural form of interaction is to update *who* or *what* appears in the scene by providing new visual references, i.e., *reference-guided subject switching*. This requires a streaming generator to immediately adopt newly specified subjects while preserving the surrounding scene and temporal dynamics accumulated before the update. Therefore, the key challenge of reference-guided subject switching lies in reconciling newly specified subjects with a causal history that entangles scene-level context with preceding-subject information. Figure 1 illustrates generation with a fixed reference and successive reference updates.

Existing methods provide complementary foundations for this interaction, yet neither directly addresses the challenge posed by dynamic reference updates in causal generation. Subject-conditioned models (Jiang et al., 2024; Wei et al., 2024; Wang et al., 2025) such as VACE and Phantom achieve strong appearance fidelity for a given set of visual references, but assume that these references remain fixed throughout generation (Jiang et al., 2025; Liu et al., 2025). Causal streaming systems, on the other hand, enable condition updates during generation; LongLive, for example, supports prompt changes by recomputing historical key/value (KV) states under updated text (Yang et al., 2025). However, subject switching fundamentally differs from such condition updates: **changing the visual subject makes the accumulated history partially incompatible with the new reference, while that same history remains essential for preserving scene and temporal continuity.** Consequently, dynamic reference updates pose a unique history-management challenge: the generator must resolve subject-specific conflicts in its causal history without sacrificing the contextual continuity carried by that history.

History-management mechanisms developed for existing causal streaming systems do not directly transfer to reference-guided subject switching. As shown in Figure 5, retaining the historical KV cache on the same adapted causal generator results in weak alignment with the updated reference, while applying the text-oriented KV recaching strategy designed for prompt switching yields similar behavior. At the other extreme, clearing the historical cache improves responsiveness to the new reference over subsequent chunks, but substantially degrades aesthetics and motion smoothness. Figure 2 summarizes this tension between adapting to the new subject and preserving temporal continuity. Resolving this tension therefore calls for semantically aware history management that differentiates subject-specific information from contextual information within the accumulated causal history.

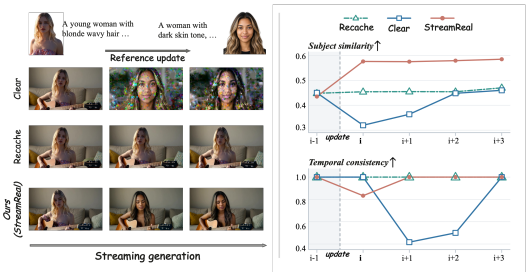


Figure 2: **Subject switching under different history strategies.** The right-hand curves report chunk-wise scores for the case shown on the left, with the text prompt and reference image switched at chunk i. Clear discards the pre-switch target KV cache; Recache recomputes historical KV states under updated text. Subject similarity and temporal consistency denote CLIP-I-Seg and Motion Smoothness, respectively. All subject-similarity scores use the updated reference, including at $i-1$, which was generated under the original conditions. Recache's nearly unchanged similarity and retained appearance indicate a failure to adopt the new subject.

Our key idea is to organize the causal history according to its semantic roles throughout streaming generation. Instead of treating accumulated history as an undifferentiated context, we continuously establish the association between historical information and the evolving subject, distinguishing subject-dependent evidence from context that remains relevant across reference updates. When the subject changes, this semantic organization allows the influence of the preceding subject to be suppressed while preserving the historical context necessary for coherent continuation.

Technically, we introduce **StreamReal**, the first framework for reference-guided subject switching in causal streaming video generation. StreamReal follows a progressive pipeline that first establishes causal reference conditioning through ***Few-Step Conditioning Alignment (FCA)*** followed by reference-conditioned Distribution Matching Distillation (DMD) warm-up, then organizes the resulting streaming history with ***Memory Flow (MF)***, and finally adapts history access to reference updates through ***Memory-Guided History Routing (MHR)***. Specifically, FCA aligns a pretrained reference pathway with the frozen few-step causal generator, and the subsequent DMD warm-up adapts the reference branch to self-generated histories, providing a reference-aware causal genera-

tor for subsequent streaming. Building on its autoregressive rollouts, MF continuously maintains a cross-chunk semantic index that associates the accumulated history with the evolving subject. When a reference update occurs, MHR consumes this semantic index to regulate access to the native historical states, suppressing history associated with the preceding subject while preserving contextual information required for coherent continuation. This progressive design integrates reference adaptation, semantic history modeling, and reference-dependent history routing, enabling responsive subject switching while maintaining scene and temporal continuity. We introduce a comprehensive Subject Switch Benchmark spanning diverse subject categories and switching positions. Experiments on OpenS2V-Eval and our benchmark demonstrate StreamReal's excellent subject fidelity and visual quality, together with a superior balance between subject-switching responsiveness and temporal continuity.

2 Related Work

Subject-conditioned video generation. Subject-conditioned video generation aims to preserve the appearance of reference subjects. Movie Weaver (Liang et al., 2025) binds prompt entities to reference images, while ConceptMaster (Huang et al., 2025b) separates concept representations to reduce identity mixing. Phantom (Liu et al., 2025) aligns textual and visual conditions, and VACE (Jiang et al., 2025) unifies reference guidance with video generation and editing. These works primarily study subject fidelity and composition for a given set of references.

Causal video generation and history management. CausVid (Yin et al., 2025) distills a bidirectional model into a few-step causal generator. Self Forcing (Huang et al., 2025a) trains on self-generated histories to reduce the training–inference mismatch, while Causal Forcing (Zhu et al., 2026) uses an autoregressive teacher for student initialization. For scene consistency, WorldKV (Yi et al., 2026) retrieves and compresses historical context to preserve previously observed scenes.

Interactive streaming control. LongLive (Yang et al., 2025) supports prompt changes by recomputing cached states under updated text. MotionStream (Shin et al., 2026) enables trajectory and camera control, while StreamAvatar (Sun et al., 2026) combines audio-driven interaction with persistent reference conditioning. StreamReal focuses on reference-image updates during causal generation, combining adaptation of reference conditioning with selective history routing to support subject changes while retaining context useful for continuation.

3 Method

3.1 Overview

To preserve the pretrained causal generator's temporal continuity while supporting reference conditioning and updates during streaming generation, we introduce ***StreamReal***, which comprises a reference-conditioning adaptation pipeline and an interaction mechanism for subject-aware memory control. First, to enable the streaming video generator to condition its generation on reference images, we perform **Few-Step Conditioning Alignment (FCA)**, followed by reference-conditioned Distribution Matching Distillation (DMD) to warm up the reference branch before memory flow training (Section 3.2). Furthermore, we found that interacting with the reference signal can leave cached history associated with the previous subject. To address this limitation, we introduce **Memory Flow (MF)**, which maintains a cross-chunk semantic index by reading historical memory summaries, labeling current tokens, and writing subject/background summaries for subsequent chunks. We further propose **Memory-Guided History Routing (MHR)**, which uses the stored token labels at each reference update to control access to native historical keys, suppressing history tied to the superseded subject while preserving context useful for temporal continuity (Section 3.3).

3.2 Reference Conditioning for a Causal Generator

We first establish reference conditioning on a frozen causal video generator (Yang et al., 2025; Zhu et al., 2026), while the reference branch is initialized from a bidirectionally pretrained VACE adapter (Jiang et al., 2025). Following the conditioning design of VACE, given a text prompt y and

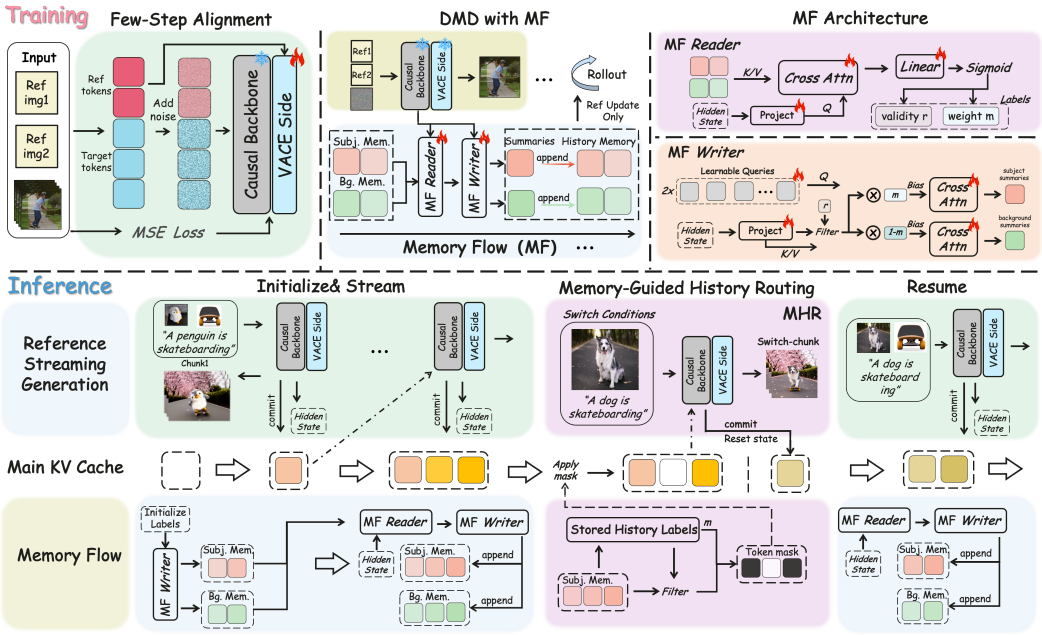


Figure 3: **Overview of StreamReal. Training (top):** FCA followed by DMD warm-up adapts the reference branch to a frozen causal backbone; Memory Flow (MF) then learns from token labels with both branches frozen. **Inference (bottom):** MF reads subject/background summaries, labels committed tokens, and writes summaries for subsequent chunks. At reference updates, Memory-Guided History Routing (MHR) uses stored labels to mask attention to old-subject keys in the native KV cache for the first updated chunk. Obsolete history is then retired and MF reseeded. Snowflakes and flames denote frozen and trainable modules, respectively.

reference images $\mathcal{R} = \{I_i\}_{i=1}^{N_r}$, we independently encode each reference image into a single latent frame and prepend the resulting reference latent r to the target video latents x along the temporal dimension, forming $z = [r; x]$. The full sequence z is then noised and fed into the main branch. In parallel, the reference branch takes the clean reference latents r as input and injects their projected features into each layer of the backbone. After denoising, only the target latents are retained and decoded into the output video, while the reference prefix is discarded.

Few-Step Conditioning Alignment (FCA). To enable the multi-step bidirectional conditioning network to operate under few-step inference while preserving effective reference consistency, we first pre-align its reference injection with the frozen few-step video generator. The target videos are synthesized by VACE-14B (Jiang et al., 2025), which also serves as the teacher model for subsequent DMD. During training, we partition the target latents x into chunks of three consecutive latent frames. In both branches, target tokens can attend to all reference tokens and to target tokens in their current and preceding chunks, whereas reference tokens attend only to one another. This preserves bidirectional attention within each chunk while excluding future target information. For the clean sequence $z = [r; x]$, we sample Gaussian noise ϵ and construct the noisy input as:

$$z_\sigma = (1 - \sigma)z + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where the noise level σ is sampled from the generator's four-step denoising schedule and remains constant within each target chunk. Let θ and ϕ denote the parameters of the main and reference branches, respectively, and $v_{\theta,\phi}$ the resulting velocity predictor. Prediction and target are:

$$\hat{v} = v_{\theta,\phi}(z_\sigma; \sigma, y, r), \qquad v^* = \epsilon - z. \tag{2}$$

Although excluded from the output video, the noised reference tokens remain in the denoising sequence and are visible to all target chunks. Since a target-only objective leaves their denoising predictions unsupervised, we additionally apply a reference loss to provide denoising supervision for the reference tokens. The parameters ϕ are optimized as:

$$\mathcal{L}_{\text{FCA}} = \mathbb{E}_{z,\epsilon,\sigma}\left[\|\hat{v}_x - v_x^*\|_2^2 + \lambda_{\text{ref}}\|\hat{v}_r - v_r^*\|_2^2\right], \tag{3}$$

where λ_{ref} controls the strength of auxiliary reference supervision. This few-step pre-alignment provides a supervised initialization for the subsequent DMD warm-up.

DMD Warm-up for Reference-Conditioned Streaming Generation. FCA pre-aligns reference injection on paired sequences, whereas streaming inference conditions on self-generated history. To mitigate accumulated distribution shift during continuous streaming, we adapt the reference branch with DMD (Yin et al., 2024) using autoregressive self-rollouts (Huang et al., 2026), while matching the denoising schedule and cache policy used during inference. Given an autoregressive rollout, we sample a noise level σ and perturb the full sequence using Eq. equation 1, sharing the same σ across all frames. DMD adapts the reference branch by matching the generated distribution to a frozen real data model through the difference between their denoising predictions:

$$\nabla_\phi \mathcal{L}_{\text{DMD}} = \mathbb{E}_{\hat{z},\,\sigma}[w(\sigma)\,(\epsilon_{\text{real}}(\hat{z}_\sigma, \sigma, c) - \epsilon_{\text{fake}}(\hat{z}_\sigma, \sigma, c))\,\nabla_\phi \hat{x}_\phi]\,. \tag{4}$$

Here, ϵ_{real} and ϵ_{fake} denote the frozen real and fake denoising predictions, $c = (y, r)$ the text-reference condition, and $w(\sigma)$ the DMD weighting. Both models evaluate the same noisy rollout $\hat{z}_\sigma$ at the same σ. To preserve reference sensitivity, we additionally apply reference guidance alongside standard text CFG only to ϵ_{real}. The resulting gap between the real and fake predictions provides distribution level supervision under self-generated history.

3.3 Memory Flow Routing for Reference Updates

We initially explored controlling subject information in the rollout history by modifying the way reference was injected. However, we observe that *the current chunk relies heavily on subject cues already present in the history* , preventing the intended subject transition. To address this limitation, we introduce **Memory Flow (MF)**, which maintains a chunk-wise read/write process over subject/background memory streams: it reads visible historical summaries, labels newly committed tokens, and writes group-specific summaries for subsequent chunks. The memory flow provides a semantic index of the rollout history without replacing the native key/value (KV) caches. We further propose **Memory-Guided History Routing (MHR)**, which uses the stored token labels at each reference update to control access to native historical keys, suppressing stale subject-specific information while preserving context useful for temporal continuity.

Memory Flow. MF first reads the visible subject/background memory summaries to label newly committed tokens, and then writes group-specific summaries for the current chunk into the corresponding memory streams. For chunk k, let H_k denote its detached token features, Pos_k their positional encodings, and $\mathcal{S}_{<k}$ the subject and background memory summaries from earlier visible chunks. The reader $\mathcal{Q}_{\psi_r}$, with trainable parameters ψ_r, uses these inputs to predict two scores for each current token:

$$(\hat{r}_k, \hat{m}_k) = \text{sigmoid}\,(\mathcal{Q}_{\psi_r}(H_k, \mathcal{S}_{<k}, \text{Pos}_k))\,, \tag{5}$$

where the sigmoid function maps each logit to a score in $[0, 1]$. For each token, $\hat{r}_k$ estimates how much of the generator's attention goes to previously labeled history, while $\hat{m}_k$ estimates what fraction of that attention is directed to the subject. Both outputs provide a token-wise score.

At inference, we use the reader predictions (r_k, m_k) as token labels. Before writing the memory summaries, we use r_k to filter out tokens with low relevance to the labeled history, thereby reducing interference from irrelevant tokens in the resulting summaries. For each valid token i, we assign subject and background writing weights $m_{k,i}$ and $1 - m_{k,i}$, respectively. Each group $g \in \{\text{sub}, \text{bg}\}$ has a separate writer that summarizes H_k using the corresponding weight vector w_k^g and positional encodings Pos_k:

$$S_k^g = \mathcal{W}_{\psi_w}^g(H_k, w_k^g, \text{Pos}_k), \tag{6}$$

where $\mathcal{W}_{\psi_w}^g$ denotes the group-specific writer and ψ_w denotes the trainable writer parameters. The weight vector w_k^g assigns one weight to each token, and the resulting group summary $S_k^g \in \mathbb{R}^{n \times d}$ becomes available to later reader calls, where n and d denote the number of memory tokens and the summary dimension, respectively. MF therefore maintains the cross-chunk memory index used to label subsequent chunks, whereas MHR uses the stored token labels to determine which native KV entries remain accessible when the reference is updated.

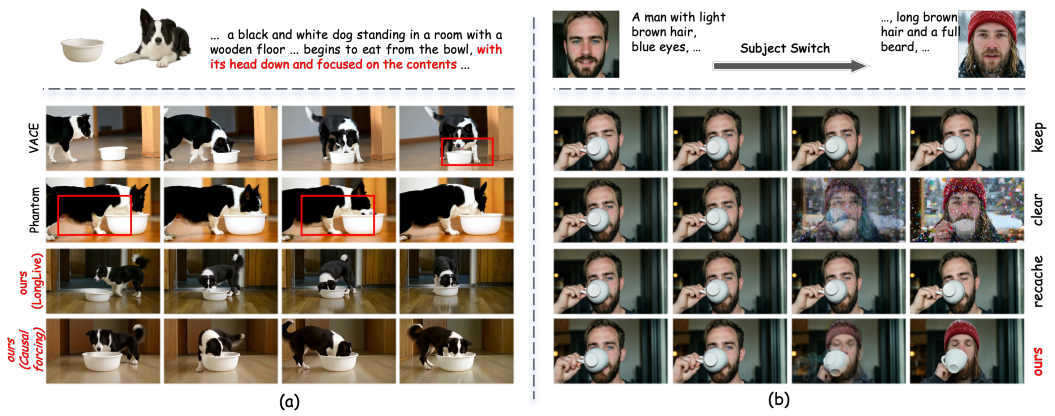


Figure 4: **Qualitative comparisons of S2V generation and subject switching.** (a) A dog-and-bowl example compares VACE, Phantom, and our LongLive and Causal Forcing variants; red boxes mark selected baseline details. (b) Subject switching with Keep, Clear, Recache, and StreamReal on our LongLive-based generator. References and prompt excerpts appear above each panel; frames run left to right within each row.

Reading and Writing the Memory Streams Each operator uses a single cross-attention layer. The reader $\mathcal{Q}_{\psi_r}$ uses current tokens as queries and the visible historical memory summaries as keys and values. With reader-specific linear projections f_r and p_r, it predicts the token labels as

$$(\hat{r}_k, \hat{m}_k) = \text{sigmoid}\left(\text{Linear}\left(\text{Attn}\left(f_r(H_k) + p_r(\text{Pos}_k), \text{Enc}(\mathcal{S}_{<k})\right)\right)\right), \tag{7}$$

where Enc concatenates the visible historical subject and background summaries and adds MLP encodings of their group statistics and relative time. As in the writer, Attn uses its first input as queries and its second as keys and values. The linear head produces two logits per token, followed by the sigmoid in Eq. equation 5. Thus, the writer compresses token information into the memory streams, whereas the reader queries those streams to label current tokens.

For each group g, the writer $\mathcal{W}^g_{\psi_w}$ uses learnable summarization queries U^g to attend to the committed tokens:

$$S^g_k = \text{Attn}_g(U^g, f_g(H_k) + p_g(\text{Pos}_k); w^g_k) + b^g_k, \tag{8}$$

where f_g and p_g are linear projections of token features and positional encodings, respectively. Their sum serves as both keys and values, while U^g provides the queries. The writing weights w^g_k control token contributions through a logarithmic attention bias. The term b^g_k is an MLP encoding of group statistics, broadcast to all summary tokens. The subject and background writers use separate parameters, each aggregating the chunk's tokens into a fixed-length group memory summary.

DMD with MF. After the DMD warm-up, we freeze both the main and reference branches (θ, ϕ) and train MF on autoregressive rollouts from the adapted generator. A frozen teacher provides token labels for each committed chunk to supervise the reader predictions in Eq. equation 5 and form the writing weights in Eq. equation 6. The writer receives gradients from this token-label loss through its summaries S^g_k when they are consumed by the reader in the subsequent chunk, while the input features H_k remain detached. The reader and writer parameters (ψ_r, ψ_w) are thus jointly optimized by token-label supervision; DMD supplies the adapted rollout distribution, rather than the loss used to update MF.

Memory-Guided History Routing. MHR uses the stored token labels from MF to control access to historical subject information when the reference is updated. Let $\mathcal{H}_{\text{old}}$ denote the set of token indices in the native history preceding the update. MHR marks valid subject tokens in this history as:

$$\mu^{\text{sub}}_i = \mathbf{1}[i \in \mathcal{H}_{\text{old}}]\, e_i\, \mathbf{1}[m_i \geq \tau_m], \tag{9}$$

where m_i denotes the stored subject label with the chunk index omitted, $e_i \in \{0, 1\}$ indicates whether the label is valid, and τ_m is the threshold for subject membership. During denoising of the

Table 1: **S2V and T2V generation results.** (a) S2V generation on OpenS2V-Eval, with latency measured in seconds per frame. (b) T2V generation on VBench; † denotes scores reused from the corresponding backbones for our variants. Total, Quality, and Semantic scores in (b) are reported on a 0–100 scale; the four individual metrics use a 0–1 scale. Throughout this section, the best and second-best values are highlighted separately for each metric within each table or panel.

(a) S2V generation

		Aesthetics	*Motion*		*Text Consistency*	*Subject Fidelity*				
Method	**Latency ↓**	**AES ↑**	**MS ↑**	**MA ↑**	**GME ↑**	**CLIP-I ↑**	**DINO-I ↑**	**CLIP-I-Seg ↑**	**DINO-I-Seg ↑**	**Nexus ↑**
VACE	2.096	0.4439	0.9715	0.1731	0.7095	0.6323	0.4005	0.6540	0.4576	0.3961
Phantom	1.955	0.4259	0.9282	0.1709	0.6964	0.6542	0.4129	0.6712	0.4662	0.4030
Ours (LongLive)	0.184	0.4575	0.9892	0.4006	0.7277	0.6086	0.3546	0.6266	0.4038	0.4052
Ours (Causal Forcing)	0.180	0.4667	0.9925	0.3135	0.6937	0.6586	0.4624	0.6729	0.5086	0.4243

(b) T2V generation

	Overall Scores			*Temporal Consistency and Dynamics*			
Method	**Total Score ↑**	**Quality Score ↑**	**Semantic Score ↑**	**Subject Consistency ↑**	**Background Consistency ↑**	**Motion Smoothness ↑**	**Dynamic Degree ↑**
CausalVid	80.7406	83.6575	69.0726	0.9644	0.9569	0.9807	0.6250
Self-Forcing	82.0308	84.7395	71.1956	0.9624	0.9588	0.9865	0.6750
LongLive / **Ours (LongLive)**†	81.3876	84.0620	70.6898	0.9686	0.9615	0.9828	0.6056
Causal Forcing / **Ours (Causal Forcing)**†	82.6153	85.5168	71.0093	0.9672	0.9555	0.9820	0.7722

first chunk after the reference update, MHR excludes keys with $\mu_i^{\text{sub}} = 1$ from attention over the main history. After denoising this chunk, we retire obsolete target KV entries together with their associated token labels and memory summaries.

To initialize MF for the new reference stage, we use Grounding DINO and SAM2 (Liu et al., 2024; Jiaxing & Hao, 2025) to generate seed labels for the subject region from the first chunk. Seed generation runs in parallel with denoising of the subsequent chunk, so the initialized memory summaries are ready before the reader processes that chunk without incurring additional latency. MHR remains inactive during ordinary generation, preserving the generator's native causal behavior.

4 Experiments

4.1 Experimental Setup

Benchmarks and baselines. We evaluate StreamReal variants built on LongLive and Causal Forcing against VACE (Jiang et al., 2025) and Phantom (Liu et al., 2025) on OpenS2V-Eval (Yuan et al., 2026) (90 single-subject and 90 multi-subject cases) using a shared evaluation protocol. We also report T2V results on VBench (Huang et al., 2024) (Appendix A.1).

Subject switching. Our Subject Switch Benchmark comprises 162 person and accessory switching cases spanning three switching positions. All switching strategies use the same LongLive-based generator and receive updated text prompts and reference images at the switching chunk. The baselines disable MF: Keep retains the historical target KV cache, Clear discards it entirely, and Recache follows LongLive's KV recaching strategy (Yang et al., 2025). Table 2 defines the score changes.

Metrics. Following the OpenS2V protocol, we measure Aesthetics (AES), Motion Smoothness (MS), Motion Amplitude (MA), text alignment (GME), and subject fidelity (CLIP-I, DINO-I, their "-Seg" variants, and Nexus). The "-Seg" variants evaluate subject regions extracted with Grounded-

Table 2: **Comparison of history strategies for subject switching.** All strategies use our LongLive-based generator. For each metric m, we report $\Delta = m_{k+1} - m_{k-1}$, excluding switching chunk k. Text- and reference-alignment metrics are evaluated against the original conditions at $k-1$ and the updated conditions at $k+1$.

	Quality	*Motion*		*Text Consistency*	*Multi-subject similarity*			
Variant	**AES** $\Delta\uparrow$	**MS** $\Delta\uparrow$	**MA** $\Delta\uparrow$	**GME** $\Delta\uparrow$	**CLIP-I** $\Delta\uparrow$	**CLIP-I-Seg** $\Delta\uparrow$	**DINO-I** $\Delta\uparrow$	**DINO-I-Seg** $\Delta\uparrow$
Keep	−0.0319	−0.0051	0.0356	−0.1175	−0.2746	−0.2895	−0.2333	−0.2688
Recache	−0.0584	−0.0134	0.0428	−0.1212	−0.2800	−0.2948	−0.2402	−0.2740
Clear	−0.3518	**−0.3729**	0.0925	−0.0927	−0.1373	−0.1666	−0.1585	−0.1889
Ours	0.0749	−0.0103	−0.0407	−0.0088	−0.0858	−0.0941	−0.1252	−0.1270

SAM (Jiaxing & Hao, 2025). Training details, baseline configurations, benchmark construction, and metric definitions are provided in Appendix A.1.

4.2 Subject-to-Video Results

Qualitative results. Figure 4(a) illustrates generation conditioned on separate dog and bowl references. Both StreamReal variants reproduce the dog's black-and-white markings and the white bowl, with the dog's head lowered toward the bowl in the later frames. These visual details illustrate reference fidelity alongside the interaction specified by the prompt.

Quantitative results. Table 1(a) shows that our Causal Forcing variant achieves the highest scores on all five subject-fidelity metrics among the compared S2V methods. Relative to Phantom, it improves DINO-I from 0.4129 to 0.4624 and DINO-I-Seg from 0.4662 to 0.5086, while also achieving the highest aesthetics and motion smoothness. Our LongLive variant achieves the highest GME and motion amplitude. Both variants achieve 0.180–0.184 s/frame, corresponding to a speedup exceeding $10\times$ relative to VACE and Phantom in the reported latency comparison. Table 1(b) reports T2V performance using the corresponding backbone scores.

4.3 Subject Switching Results

Qualitative results. Figure 4(b) compares subject switching to a reference with long hair and a red hat. StreamReal incorporates both appearance changes while retaining the indoor setting and cup-drinking pose in the displayed comparison. Keep and Recache retain the original appearance, whereas Clear introduces the updated appearance together with pronounced block artifacts across the subject and background.

Quantitative results. Table 2 shows that StreamReal achieves the largest aesthetics gain and the smallest adjacent-chunk decreases in GME and all four reference-similarity metrics. Its GME change is -0.0088, compared with -0.1175 for Keep and -0.1212 for Recache. Clear's motion-amplitude increase of 0.0925 accompanies a motion-smoothness decrease of 0.3729, whereas StreamReal limits the smoothness decrease to 0.0103. These adjacent-chunk results indicate effective adaptation to updated conditions with limited disruption to motion smoothness.

Chunk-wise switching response. Figure 5 tracks adaptation to the updated conditions at the chunk level. StreamReal's GME, CLIP-I-Seg, and DINO-I-Seg increase sharply at the first updated chunk i and remain the highest among the compared strategies through $i+3$. Keep and Recache maintain high motion smoothness but show limited improvement in alignment with the new conditions. Clear's subject alignment improves gradually, while its text alignment begins to recover at $i+1$.

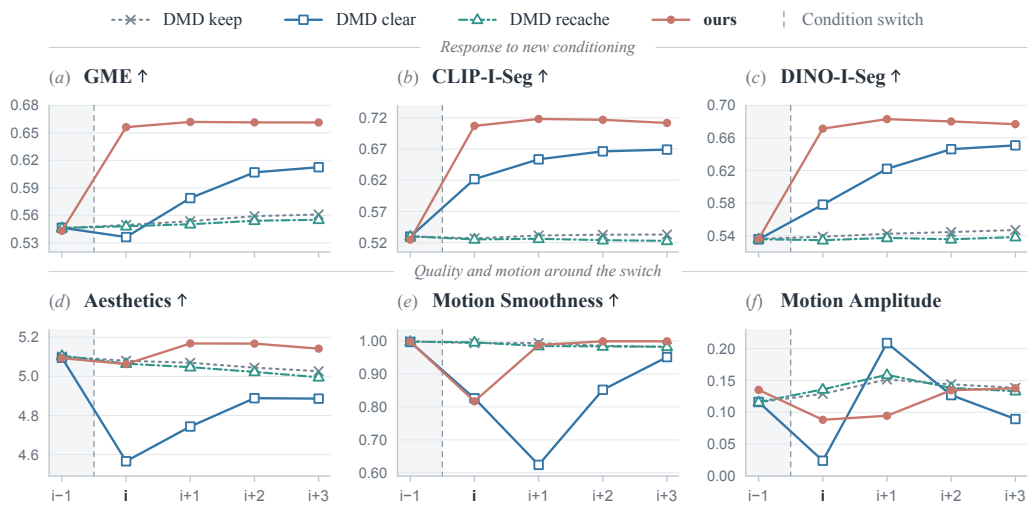


Figure 5: **Chunk-wise response to a subject switch.** Each point averages the score for a given chunk over the same 162 switching cases. The dashed line marks the condition update immediately before chunk i, the first chunk generated under the updated conditions. Shading identifies chunk $i-1$, generated under the original conditions. The text- and reference-alignment metrics in the top row use the updated prompt and reference image for all displayed chunks, including $i-1$.

Table 3: **Ablation study of video quality and subject fidelity.**

	Aesthetics	*Motion*		*Text Consistency*	*Subject Fidelity*				
Method	**AES ↑**	**MS ↑**	**MA ↑**	**GME ↑**	**CLIP-I ↑**	**DINO-I ↑**	**CLIP-I-Seg ↑**	**DINO-I-Seg ↑**	**Nexus ↑**
+warm DMD	0.4388	0.9642	0.3142	0.7202	0.6168	0.3885	0.6300	0.4292	0.3823
+FCA	0.4349	0.9229	0.0340	0.6823	0.6248	0.4164	0.6411	0.4400	0.4296
Ours	0.4667	0.9925	0.3135	0.6937	0.6586	0.4624	0.6729	0.5086	0.4243

4.4 Ablation Study

Quantitative results. Table 3 compares reference fidelity and motion across the evaluated variants. The full model achieves the highest aesthetics, motion smoothness, and all four CLIP/DINO similarity scores. Compared with +warm DMD, it improves DINO-I and DINO-I-Seg, with nearly unchanged motion amplitude (0.3135 vs. 0.3142). Relative to +FCA, it increases motion amplitude from 0.0340 to 0.3135 while improving all four CLIP/DINO similarity scores.

5 Conclusion

We presented StreamReal, which enables reference-guided subject switching in causal streaming video generation through reference adaptation and semantic history control. FCA aligns a pre-trained reference branch with a frozen few-step generator, and reference-conditioned DMD adapts it to self-generated histories. Memory Flow maintains a cross-chunk semantic index that distinguishes subject-associated evidence from background context. At reference updates, Memory-Guided History Routing uses this index to suppress preceding-subject evidence while retaining contextual history for the transition. On OpenS2V-Eval, the Causal Forcing variant leads all five subject-fidelity metrics among the evaluated methods. On the Subject Switch Benchmark, the LongLive variant achieves higher post-switch text and subject alignment than the three history baselines and better preserves motion smoothness than cache clearing.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

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*, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

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, 2026.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. ConceptMaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025b. URL https://arxiv.org/abs/2501.04698.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. VideoBooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6689–6700, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jiang_VideoBooth_Diffusion-based_Video_Generation_with_Image_Prompts_CVPR_2024_paper.html.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Zhang Jiaxing and Tang Hao. Sam2 for image and video segmentation: A comprehensive survey. *arXiv preprint arXiv:2503.12781*, 2025.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. *arXiv preprint arXiv:2502.07802*, 2025. URL https://arxiv.org/abs/2502.07802.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Liu_Phantom_Subject-Consistent_Video_Generation_via_Cross-Modal_Alignment_ICCV_2025_paper.pdf.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. MotionStream: Real-time video generation with interactive motion controls. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.01266.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. StreamAvatar: Streaming diffusion models for real-time interactive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.22065.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Wenchuan Wang, Mengqi Huang, Yijing Tu, and Zhendong Mao. DualReal: Adaptive joint training for lossless identity-motion fusion in video customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16565–16575, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_DualReal_Adaptive_Joint_Training_for_Lossless_Identity-Motion_Fusion_in_Video_ICCV_2025_paper.html.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. DreamVideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6537–6549, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wei_DreamVideo_Composing_Your_Dream_Videos_with_Customized_Subject_and_Motion_CVPR_2024_paper.html.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

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*, 2025.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. WorldKV: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026. URL https://arxiv.org/abs/2605.22718.

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

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*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.pdf.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Chongyang Ma, Jiebo Luo, Li Yuan, et al. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *Advances in Neural Information Processing Systems*, 38, 2026.

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.