EVERSTYLE: LONG-HORIZON CONSISTENT STREAMING VIDEO STYLIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video stylization continuously transforms input video frames into user-specified visual styles, supporting real-time applications such as live broadcasting and interactive content creation. Recent diffusion-based methods have made substantial progress toward efficient and temporally coherent streaming stylization, yet maintaining consistent stylization throughout an open-ended rollout remains challenging. In particular, previously seen content is usually stylized differently when it reappears after leaving the active context, despite smooth local segments. To address these, we propose **EverStyle**, a novel long-horizon consistent streaming video stylization framework. To recover the established appearance of recurring content, we design Source-Indexed Stylization Memory (SISM), which stores and retrieves source-indexed stylization residuals, and Replay-Based Stylization Consistency Regularization (RSCR), which trains the generator to reproduce earlier stylization under different temporal contexts. Yet reproducing earlier appearance alone, without constraining new content to the same distribution, cannot keep the evolving stylization temporally consistent, leading to global drift. To suppress this, we introduce Early-Anchored Stylization Distribution Matching (EA-SDM), which jointly evaluates early and current generations to align with the same style without content correspondence. Experiments demonstrate that EverStyle maintains competitive stylization quality, while substantially improving long-horizon consistency under both natural scene revisitation and exact content recurrence settings, sustaining consistent stylization over sequences exceeding 130,000 frames.

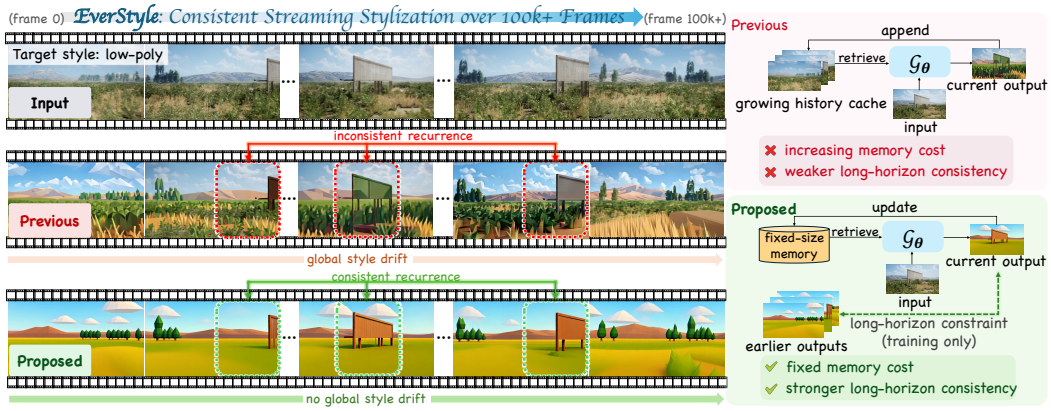

Figure 1: **Long-horizon consistent streaming video stylization using proposed EverStyle.** Existing approaches (Xiao et al., 2026; Lee et al., 2026) may restylize recurring content inconsistently or drift over long rollouts (left), or rely on a growing history cache for retrieval (right), while EverStyle adopts a fixed-size memory and constrains current generations with earlier outputs during training.

1 INTRODUCTION

Generative video editing modifies video appearance, semantics, or motion under textual or visual conditions while preserving unrelated content (Molad et al., 2023; Jiang et al., 2025). Video styliza-

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.

tion is an important instance of this task, applying a specified style while preserving source structure and motion (Song et al., 2025). Diffusion-based methods (Li et al., 2026b; Huang et al., 2024; Ouyang et al., 2024; Liang et al., 2024; Qin et al., 2024) have substantially improved editing and stylization in visual quality, temporal consistency, and controllability, but many process prerecorded videos non-causally with access to future frames. Real-time applications, including live broadcasting, interactive content creation, and visual effects, instead require these transformations to keep pace with incoming frames. Recent methods therefore enable streaming through causal generation, few-step distillation, rolling caches, and recurrent state propagation (Xing et al., 2024; Liang et al., 2025; Chen et al., 2025; Zhao et al., 2026b; Wang et al., 2026; Lyu et al., 2026). However, streaming extends the consistency requirement from short clips to the entire open-ended session, where frames must be generated causally and the stream never ends. In practice, existing methods can still render previously seen content differently when it reappears after leaving the active context (see Fig. 1). These errors can occur despite plausible local chunks and cannot be corrected after frames are emitted, thus making long-horizon consistency a critical requirement for streaming video stylization.

One key challenge is *historical stylization-state forgetting*: source content reappears beyond the local context but its earlier stylization cannot be reliably recovered. A returning subject may acquire different colors or textures. Many existing approaches preserve history through bounded windows, fixed anchors, recurrent states, feature banks, and compressed caches (Zhang et al., 2026; Ji et al., 2026; Cai et al., 2026). Such mechanisms can support local continuity and stabilize long rollouts, but eviction or compression might remove details needed to recover earlier stylizations. More recent memory-based methods explicitly retrieve past observations (Yu et al., 2025a; 2026; Lee et al., 2026), including previous target outputs matched by source content (Lee et al., 2026). This improves access to distant history, but retrieving it does not ensure recurring content recovers its earlier stylization: the generator may still produce a locally plausible stylization that differs from earlier outputs.

To address these limitations, we propose **EverStyle**, a novel streaming video stylization framework for long-horizon consistency. We first develop Source-Indexed Stylization Memory (SISM) to make previously established stylizations accessible when source content reappears. Specifically, SISM indexes historical stylization residuals by source features, allowing recurring content to retrieve how it was stylized before under a fixed memory budget. However, retrieving the historical state alone does not ensure that it can be faithfully reproduced under a changed temporal context and fresh generation noise. We therefore introduce Replay-Based Stylization Consistency Regularization (RSCR), which replays historical source content later in the whole rollout and explicitly trains the generator to recover its earlier stylized appearance. In this way, SISM and RSCR together make historical stylization both accessible and reproducible across distant recurrences. Yet the recurrence consistency alone is insufficient for open-ended video streams: both 2 mechanisms operate only when previous content reappears, leaving newly encountered content unconstrained by the style established earlier and allowing the global stylization to gradually drift over time. To alleviate such a complementary failure, we further introduce Early-Anchored Stylization Distribution Matching (EA-SDM), which jointly evaluates the current generations with early stylized frames, providing a long-range constraint that keeps later outputs compatible with the established style without requiring pixel-wise correspondence between distant frames. Together, these designs enable EverStyle to maintain consistent stylization across long causal rollouts while preserving competitive source fidelity, short-term temporal coherence, and style adherence. In summary, the contributions of this paper are fourfold:

- We propose EverStyle, a streaming video stylization method that combines explicit memory of historical stylization states with 2 complementary long-range training objectives to address historical stylization forgetting and global style drift under a fixed memory budget.
- We develop SISM to store and retrieve source-indexed stylization residuals, and introduce RSCR to reinforce their reuse, enabling recurring content to recover its established stylization beyond the local context, despite changes in temporal history and generation noise.
- We introduce EA-SDM to jointly evaluate early and current stylized outputs, suppressing global style drift without requiring pixel-wise correspondence between distant frames.
- Experiments demonstrate that EverStyle substantially improves long-horizon consistency under both natural scene revisitation and content recurrence settings, remaining consistent over sequences exceeding 130,000 frames while preserving competitive stylization quality.

2 Related Work

Generative Video Editing. Existing methods edit video appearance, semantics, or motion through per-video adaptations (Bar-Tal et al., 2022; Wu et al., 2023; Molad et al., 2023), training-free feature propagations (QI et al., 2023; Geyer et al., 2024; Yang et al., 2023; Kara et al., 2024; Yang et al., 2024), or large-scale instruction-conditioned trainings (Qin et al., 2024; Zhang et al., 2024; Jiang et al., 2025; Yang et al., 2025b; Yu et al., 2025b; Yang et al., 2026b; Liang et al., 2026). Video stylization transfers text-, image-, or first-frame-specified styles while preserving source structures and motions (Huang et al., 2024; Ouyang et al., 2024; Liang et al., 2024; Song et al., 2025; Ye et al., 2025), with recent methods extending this capability to longer videos. DreamStyle (Li et al., 2026c) propagates stylization across segments through the preceding final frame, while EchoStyle (Li et al., 2026b) combines init-follow training with sliding-window inference for minute-long videos. However, these methods primarily target prerecorded videos and do not directly address real-time streaming, where future frames are unavailable and computation must remain bounded as the stream grows.

Streaming Video Generation. Recent methods enable continuous diffusion generation via progressive clip extension or sliding denoising queues (Henschel et al., 2025; Kim et al., 2024), and through causal models with per-frame noise schedules (Chen et al., 2024; Sun et al., 2025). A complementary line distills high-quality bidirectional networks into few-step causal generators, and trains them on self-generated rollouts to reduce the inference-time exposure bias (Huang et al., 2026; Zhu et al., 2026; Zhao et al., 2026a). Long-horizon approaches further improve the rollout stability through history compressions, bounded causal attentions, and extended-horizon trainings (Zhang et al., 2026; Cui et al., 2026; Yang et al., 2025a; Liu et al., 2026b). Parallel work extends these ideas to real-time source-conditioned video translation, editing, and stylization (Chen et al., 2025; Liang et al., 2025; Xing et al., 2024; Zhao et al., 2026b; Wang et al., 2026; Lyu et al., 2026). However, their bounded temporal context often limits access to distant historical information over extended video streams.

Context Management and Memory for Video Diffusion. A lot of long-horizon video generation techniques manage history through either compressed internal context or explicitly stored, retrievable information. Internal approaches preserve selected histories using anchors, recent windows, recurrent states, feature banks, or compressed frame and Key-Value (KV) representations (Xing et al., 2024; Yang et al., 2025a; Liu et al., 2026b; Li et al., 2026a; Meng et al., 2026; Zhao et al., 2026b; Chen et al., 2025; Liang et al., 2025; Zhang et al., 2026; Ji et al., 2026; Cai et al., 2026). While efficient, selection and compression limit access to detailed past observations over long streams. Other approaches make historical information explicitly retrievable: Context-as-Memory (Yu et al., 2025a) and MemLearner (Yu et al., 2026) retrieve past observations through hand-designed rules or learned query tokens, while VMem, OneStory, VideoMemory, and PermaVid store geometric views, cross-shot content, semantic descriptions, or editable scene states (Li et al., 2025; An et al., 2026; Zhou et al., 2026; Yang et al., 2026a). Most closely related to the long-horizon consistency problem we study, Memory-V2V (Lee et al., 2026) stores previous edited segments and retrieves them by source similarity, improving cross-turn consistency. However, its external cache size grows with history, and retrieval alone cannot sufficiently ensure the same editing choice recurs when content reappears much later under a different context. This motivates fixed-size memory and stronger long-horizon consistency constraints for open-ended streaming video stylization.

3 Method

3.1 Preliminaries

Streaming Video Stylization. Given a source video partitioned into a sequence of chunks $\{\mathbf{x}_t\}_{t=1}^{T}$ and a target-style prompt p, streaming video stylization produces stylized chunks $\{\mathbf{y}_t\}_{t=1}^{T}$ causally with each chunk generated from Gaussian noise $\mathbf{z}_t$, the current source, and a bounded history $\mathbf{h}_t$ as:

$$\mathbf{y}_t = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}_t, \mathbf{x}_t, p, \mathbf{h}_t), \qquad \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\mathcal{G}_{\boldsymbol{\theta}}$ denotes the generator. Following common designs (Xing et al., 2024; Huang et al., 2026), $\mathbf{h}_t = [\mathbf{h}^{\text{sink}}, \mathbf{h}_t^{\text{recent}}]$ contains some fixed initial frames and a sliding window of recent frames. This bounded context can often support short-term continuity, but most intermediate history becomes inaccessible once it leaves the recent window. To further maintain continuity across chunk boundaries, consecutive chunks share several overlapping frames. We therefore write $\mathbf{y}_t = [\mathbf{y}_t^{\text{overlap}}, \mathbf{y}_t^{\text{new}}]$, where $\mathbf{y}_t^{\text{overlap}}$ is stylized frames shared with the previous chunk and $\mathbf{y}_t^{\text{new}}$ is newly generated frames.

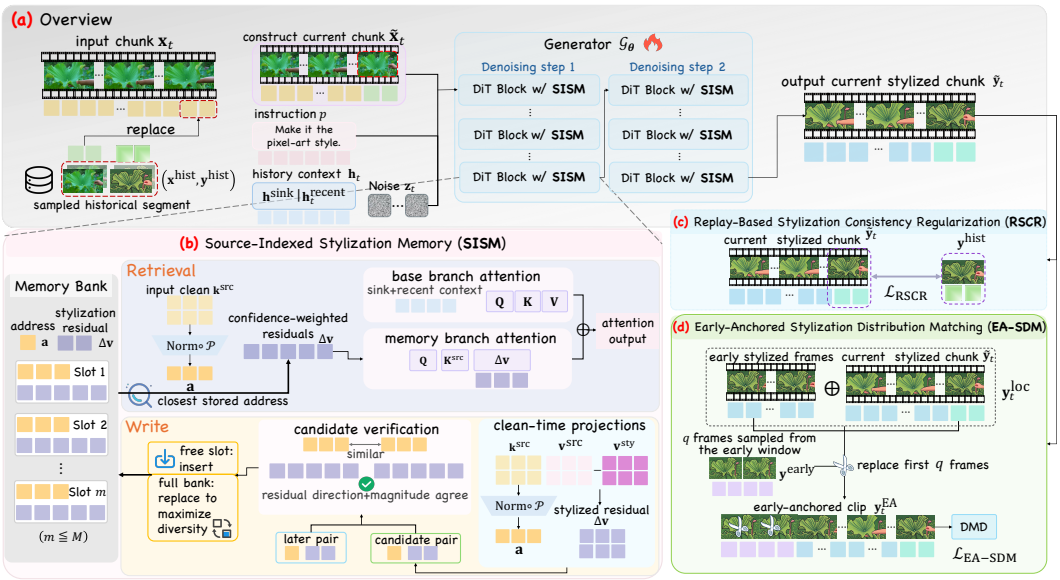

Figure 2: **Illustration of our proposed EverStyle** for long-horizon consistent streaming video stylization, which augments (a) the sink-and-recent pipeline with (b) fixed-budget, per-block SISM to retain historical stylization and (c) RSCR to train its reproduction, complemented by (d) EA-SDM to anchor global style as content changes.

Distribution Matching Distillation (DMD) (Yin et al., 2024) trains a few-step generator by matching its output distribution to the target distribution represented by a pretrained diffusion/flow model. Following its autoregressive extensions (Yang et al., 2025a), we use DMD as the local training objective for our causal generator $\mathcal{G}_{\boldsymbol{\theta}}$. For each rollout chunk, we form a local clip from the detached overlap and newly generated frames as $\mathbf{y}_t^{\text{loc}} = [\text{sg}(\mathbf{y}_t^{\text{overlap}}), \mathbf{y}_t^{\text{new}}]$. At a sampled flow timestep, we perturb this clip with noise and feed the same noisy latent to a frozen *real* flow model representing the target distribution and an online *fake* flow model fitted to the current distribution of $\mathcal{G}_{\boldsymbol{\theta}}$:

$$\mathcal{L}_{\text{DMD}} = \frac{1}{2}\left\|\mathbf{y}_t^{\text{loc}} - \text{sg}\left(\mathbf{y}_t^{\text{loc}} - \mathbf{g}_t^{\text{DMD}}\right)\right\|_{\mathbf{m}_t}^2, \quad \mathbf{g}_t^{\text{DMD}} = \frac{\widehat{\mathbf{y}}_{\text{fake}}^0 - \widehat{\mathbf{y}}_{\text{real}}^0}{\text{mean}\left|\mathbf{y}_t^{\text{loc}} - \widehat{\mathbf{y}}_{\text{real}}^0\right| + \delta}. \tag{2}$$

Here, sg/mean denote stop-gradient/average operators, $\widehat{\mathbf{y}}_{\text{real}}^0$ and $\widehat{\mathbf{y}}_{\text{fake}}^0$ are clean-latent estimates predicted by real and fake models from the same noisy version of $\mathbf{y}_t^{\text{loc}}$, and $\mathbf{g}_t^{\text{DMD}}$ denotes the normalized DMD update direction which pushes the generator toward target distribution with $\delta > 0$. The surrogate loss $\mathcal{L}_{\text{DMD}}$ is constructed so that its gradient on the optimized frames equals $\mathbf{g}_t^{\text{DMD}}$. The frame-wise binary mask $\mathbf{m}_t$ is set to zero on overlapping frames and one on newly generated frames; hence the former provide temporal context while only the latter receive direct supervision of DMD.

3.2 Source-Indexed Stylization Memory

The sink-and-recent history in Sec. 3.1 loses earlier stylization states once they leave the active context. Retaining more frames or features (Lee et al., 2026; Yu et al., 2025a; 2026) either grows memory with history or requires aggressive compression under a fixed budget. Our key hypothesis is that long-horizon stylization consistency does not require memorizing full historical stylized representations. The current source already provides most content, pose, and geometry information; when content reappears, what must be recovered is mainly *how it was previously stylized*. Our *Source-Indexed Stylization Memory* (SISM) therefore stores stylization changes relative to the source, indexed by source features, preserving past stylization within a fixed memory budget.

Source-indexed stylization representation. As illustrated in Fig. 2 (b), we instantiate SISM independently at each self-attention layer of $\mathcal{G}_{\boldsymbol{\theta}}$, since different layers operate in distinct representation spaces. Reusing the layer's existing projections under clean-time conditioning, we obtain the source key/value matrices $\mathbf{K}^{\text{src}}$, $\mathbf{V}^{\text{src}}$ and the corresponding stylized value matrix $\mathbf{V}^{\text{sty}}$. For each token and attention head, let $\mathbf{k}^{\text{src}}$, $\mathbf{v}^{\text{src}}$, and $\mathbf{v}^{\text{sty}}$ denote the corresponding row vectors, from which we define a compact source address $\mathbf{a}$ and a full-dimensional stylization residual representation $\Delta\mathbf{v}$ as

$$\mathbf{a} = \text{Norm}\left(\mathcal{P}(\mathbf{k}^{\text{src}})\right), \qquad \Delta\mathbf{v} = \mathbf{v}^{\text{sty}} - \mathbf{v}^{\text{src}}. \tag{3}$$

The fixed, parameter-free projection $\mathcal{P}$ divides channels evenly into D consecutive groups and averages each group, yielding a D-dimensional vector. Norm then applies ℓ_2 normalization for cosine matching. We use source features as addresses because stylized features may drift over long rollouts. Only addresses are compressed for efficient retrieval; $\mathbf{a}$ encodes source content, while $\Delta\mathbf{v}$ keeps the original value dimension to preserve detailed stylization changes relative to the source.

Fixed-budget memory and recall. At each self-attention layer of $\mathcal{G}_{\boldsymbol{\theta}}$, SISM stores source addresses and residuals of at most M historical frames plus one temporary candidate slot. To avoid retaining transient generation errors, a frame's addresses and residuals first enter this slot. Once a later source frame is stylized, we match their source tokens by address similarity and compare matched residuals. Verification requires enough shared source content and agreement in residual direction and magnitude; otherwise, the later frame's addresses and residuals replace the candidate. A verified candidate fills an empty slot or, if all M slots are full, replaces an entry only when this reduces the bank's maximum pairwise source similarity; otherwise, it is discarded. This fixed-size bank favors distinct source content over redundant frames. During streaming generation, each current source token retrieves the residual at the closest stored source address. A separate attention branch uses confidence-weighted residuals as values and current source keys, then adds its output to the original sink-and-recent attention. Residuals specify *how* content was stylized, while current source features guide *where* to apply the changes. Without reliable matches, zero correction leaves original path unchanged. The overall design is illustrated in Fig. 2 (b), while Appx. A provides further details on memory writing, retention, confidence-calibrated retrieval, and fusion.

3.3 Replay-Based Stylization Consistency Regularization

SISM makes an established historical stylization accessible when related source content reappears. Accessibility alone, however, does not ensure consistent stylization upon recurrence: the generator still operates under different temporal history and new noise realization, and may stylize the same source content differently even when its historical state is retrieved. Long-horizon consistency requires the established source-stylization mapping to remain stable under such rollout-state changes. We therefore introduce *Replay-Based Stylization Consistency Regularization* (RSCR), which explicitly trains $\mathcal{G}_{\boldsymbol{\theta}}$ to reproduce earlier stylization when the same content is replayed later in the rollout.

For the current source $\mathbf{x}_t$ with N frames, we sample an earlier contiguous r-frame source-stylization pair $(\mathbf{x}^{\text{hist}}, \mathbf{y}^{\text{hist}})$, where $\mathbf{y}^{\text{hist}}$ is previously generated stylization output of $\mathbf{x}^{\text{hist}}$. Using $[:]$ to denote a contiguous frame slice, our proposed RSCR loss based on source replay is defined as

$$\mathcal{L}_{\text{RSCR}} = \frac{1}{2}\left\|\widetilde{\mathbf{y}}_t[N-r+1:N] - \mathrm{sg}(\mathbf{y}^{\text{hist}})\right\|^2. \tag{4}$$

Here, $\widetilde{\mathbf{y}}_t = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}_t, \widetilde{\mathbf{x}}_t, p, \mathbf{h}_t)$ regenerates stylization under current history $\mathbf{h}_t$ and fresh noise $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, as in Sec. 3.1. Its input $\widetilde{\mathbf{x}}_t = [\mathbf{x}_t[1:N-r], \mathbf{x}^{\text{hist}}]$ replaces the last r source frames with historical content. Thus, $\widetilde{\mathbf{y}}_t[N-r+1:N]$ and $\mathbf{y}^{\text{hist}}$ stylize the same source $\mathbf{x}^{\text{hist}}$ under current and earlier rollout states. Minimizing $\mathcal{L}_{\text{RSCR}}$ penalizes changes to this established mapping, encouraging recurring content to recover its earlier appearance despite changed context and noise.

3.4 Early-Anchored Stylization Distribution Matching

SISM and RSCR reduce historical stylization-state forgetting by retrieving earlier stylizations and training their reproduction. However, both mechanisms are tied to recurring source contents and do not constrain newly appearing content to follow the established style. A rollout may thus stay locally coherent and reproduce revisited content consistently, yet still drift gradually in color, texture, or overall appearance—a failure the traditional local DMD objective (Sec. 3.1) can hardly expose, since it evaluates only a few nearby video frames. To provide a long-range reference as the global stylization state, we introduce *Early-Anchored Stylization Distribution Matching* (EA-SDM), which evaluates current generations together with stylized frames sampled from the start of the rollout.

Let L denote DMD clip length and $\mathbf{y}_t^{\text{loc}}$ the local clip defined in Sec. 3.1. Starting from the second chunk in streaming rollout, we randomly sample an anchor length $q \in \{1, \cdots, L-1\}$, and a q-frame stylized segment $\mathbf{y}^{\text{early}}$ from the earliest portion of the whole rollout. We form an early-anchored clip $\mathbf{y}_t^{\text{EA}} = [\mathrm{sg}(\mathbf{y}^{\text{early}}), \mathbf{y}_t^{\text{loc}}[q+1:L]]$ and evaluate it using the same DMD construction as in Sec. 3.1:

$$\mathcal{L}_{\text{EA-SDM}} = \frac{1}{2}\left\|\mathbf{y}_t^{\text{EA}} - \mathrm{sg}\big(\mathbf{y}_t^{\text{EA}} - \mathbf{g}_t^{\text{EA}}\big)\right\|^2_{\mathbf{m}_t^{\text{EA}}}. \tag{5}$$

Table 1: **Quantitative comparison** of stylization quality on the style-editing subsets of IVEBench (Chen et al., 2026) (left) and long-horizon consistency on Context-as-Memory and IVEBench-Cycle (right). Short and Long denote the short- and long-video partitions. The best and second-best results are marked in **bold** and underline. †: StableV2V runs out of memory on some Long-subset videos.

Method	IVEBench-Style (Short/Long)			
	Total Score ↑	Video Quality ↑	Instruction Compliance ↑	Video Fidelity ↑
StableV2V†	0.598/0.561	0.634/0.628	0.554/0.540	0.607/0.514
EchoStyle	0.584/0.598	0.644/0.629	0.472/0.505	0.634/0.659
StreamV2V	0.601/0.599	0.697/0.728	0.443/0.396	0.664/0.673
LiveEdit	0.610/0.639	**0.816/0.825**	0.214/0.223	0.800/**0.870**
SANA-Streaming	0.700/0.684	0.714/0.726	0.620/0.575	0.767/0.752
JoyAI-Video-Edit	0.739/0.744	0.774/0.786	0.630/0.643	0.814/0.803
Ours	**0.758/0.756**	0.799/0.809	**0.655/0.648**	**0.821**/0.811

Method	Context as Memory		IVEBench-Cycle	
	LPIPS ↓	DINO ↑	LPIPS ↓	DINO ↑
StreamV2V	0.4224	0.7036	0.0467	0.9741
LiveEdit	0.4609	0.6268	0.1431	0.8965
SANA-Streaming	0.5224	0.6946	0.4132	0.7618
JoyAI-Video-Edit	0.4852	0.7060	0.0865	0.9584
Ours	**0.2830**	**0.8051**	**0.0430**	**0.9786**

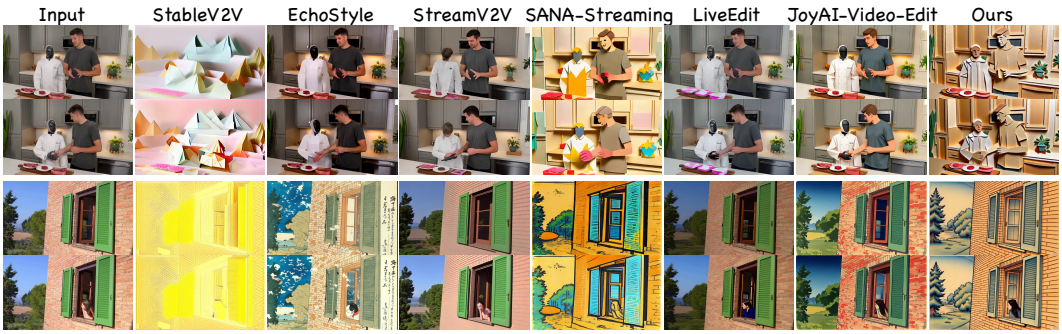

Figure 3: **Qualitative comparisons** on the style-editing subset of IVEBench. Each pair of rows corresponds to one video, shown with two representative frames. Target styles from top to bottom are: layered paper craft and Ukiyo-e. Our method expresses the requested style while better preserving subjects, scene layout, and fine structural details across frames.

Here, $\mathbf{g}_t^{\text{EA}}$ is the normalized DMD direction computed from $\mathbf{y}_t^{\text{EA}}$ following the same real/fake scoring procedure as $\mathbf{g}_t^{\text{DMD}}$, while $\mathbf{m}_t^{\text{EA}}$ is zero on the detached early frames and follows the local DMD mask on the current suffix. Hence, the earlier frames serve only as a fixed reference, whereas gradients act on the current generation. Evaluating both within the same DMD clip makes the update of current frames depend on stylization established near the beginning of the rollout, encouraging later outputs to remain compatible with such a long-range appearance reference. Unlike RSCR, EA-SDM does not require the same source input content to recur or distant frames to match pixel-wise, thereby complementing content-specific recurrence consistency with a global constraint against style drift.

Overall objective. With SISM incorporated into the forward generation path of $\mathcal{G}_{\boldsymbol{\theta}}$, we retain the original local DMD objective and jointly optimize the two long-range consistency objectives as:

$$\mathcal{L} = \mathcal{L}_{\text{DMD}} + b_{\text{rep}}\mathcal{L}_{\text{RSCR}} + \lambda_{\text{EA}}\mathcal{L}_{\text{EA-SDM}}, \qquad b_{\text{rep}} \sim \text{Bernoulli}(p_{\text{rep}}). \tag{6}$$

Here, p_{rep} controls how often RSCR is activated, while λ_{EA} controls the strength of EA-SDM. For the first rollout chunk during training, where no early reference is available, EA-SDM is omitted.

4 Experiments

4.1 Implementation Details

We implement our EverStyle by adapting LongLive (Yang et al., 2025a) to source-conditioned video stylization. We first fine-tune the diffusion Transformer (DiT) of a reference-free Wan-Fun/VideoX-Fun variant (aigc apps, 2026), built upon Wan2.2-TI2V-5B (Wan et al., 2025), with a flow-matching objective on the Global Style Transfer subset of Ditto-1M (Bai et al., 2026), then distill it into a two-step causal autoregressive generator via the recipe of LongLive. After ODE initialization (Yin et al., 2025), the student is trained with short-horizon self-forcing DMD for 5K iterations (full-parameter, with learning rate 2×10^{-6}), using three latent frames per block, sink size 3, and local attention size 15. For long-horizon tuning, we randomly sample source videos from Ditto-1M, temporally repeat them to LongLive's default long-horizon length, and sample style instructions from the Global Style Transfer pool, optimizing a rank-512 LoRA (with learning rate 7×10^{-6}) for 20K iterations. SISM, RSCR, and EA-SDM are activated only in this stage with $M = 15$, $p_{\text{rep}} = 0.5$, and $\lambda_{\text{EA}} = 1$.

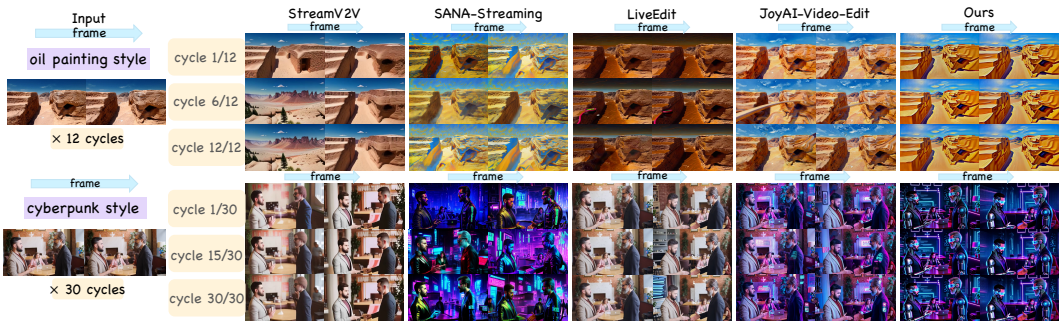

Figure 4: **Qualitative comparisons** on IVEBench-Cycle (source clip looped over 6,000 frames). For each case, rows show the first, 50%-position, and final cycles, with two frames at matching within-cycle positions. Our method preserves both the requested style and source structure throughout.

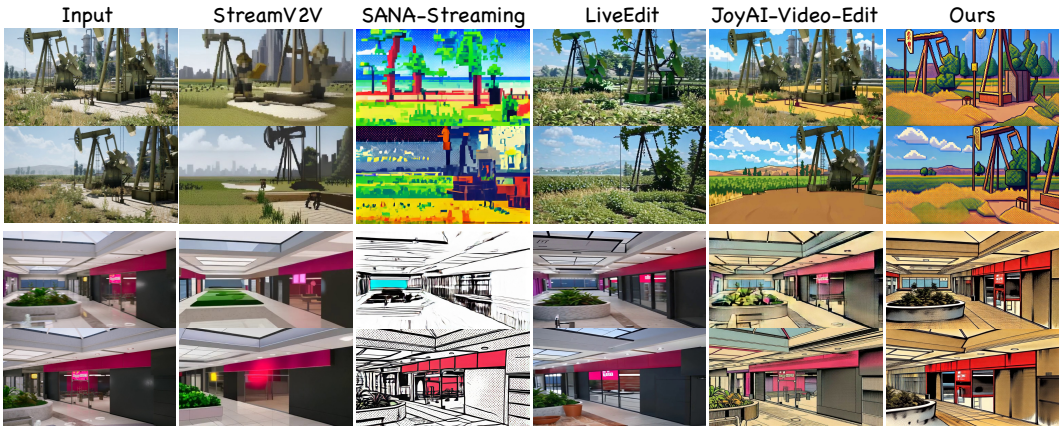

Figure 5: **Qualitative comparisons** on Context-as-Memory (CAM). Each pair of rows shows two temporally distant frames of the same scene. Target styles from top to bottom are: pixel art and vintage comic book ink. EverStyle maintains consistent appearance across revisits.

4.2 Comparison with State-of-the-Art Methods

We evaluate EverStyle from two complementary aspects: stylization quality and long-horizon consistency. For stylization quality, we follow the IVEBench style-editing protocol (Chen et al., 2026) on 143 videos (93 Short and 50 Long) and compare against six representative methods: StableV2V (Liu et al., 2026a), EchoStyle (Li et al., 2026b), StreamV2V (Liang et al., 2025), LiveEdit (Wang et al., 2026), SANA-Streaming (Zhao et al., 2026b), and JoyAI-Video-Edit (Xiao et al., 2026). We report the official Total Score, Video Quality, Instruction Compliance, and Video Fidelity metrics. While these metrics comprehensively assess stylization quality, the standard Short/Long protocol does not compare outputs across large temporal gaps and therefore cannot directly measure whether an established stylization remains consistent over a long causal rollout. We therefore evaluate long-horizon consistency on two complementary benchmarks. IVEBench-Cycle measures exact content recurrence by repeating 10 diverse IVEBench source videos into sequences exceeding 6,000 frames, such that identical content reappears across distant cycles. CAM (Yu et al., 2025a) measures natural scene revisitation using five 7,601-frame sequences with viewpoint variations and field-of-view overlap annotations. Together, the two benchmarks evaluate recurrence consistency under controlled exact repetition and natural revisitation, respectively. For both 2 benchmarks, we compare an early stylized reference with temporally distant recurrences and report LPIPS (Zhang et al., 2018) and DINO (Caron et al., 2021). On IVEBench-Cycle, we pair identical within-cycle positions between the first cycle and later cycles at 25%, 50%, 75%, and the final position; on CAM, we use the overlap annotations to pair an early reference with its distant revisits. These pairings focus the evaluation on changes in the stylization of recurring content rather than unrelated source variation. StableV2V and EchoStyle are excluded from this evaluation because they are non-causal, while all remaining methods are run on full sequences using their native streaming pipelines. Finally, we stress-test long-horizon stability on a 96.7-minute, 139,210-frame Singapore city-walk video (POPtravel, 2019).

Stylization quality. As exhibited in Tab. 1 (left), EverStyle achieves competitive stylization quality on both IVEBench splits, leading in Total Score (0.758/0.756 for Short/Long) and Instruction Compliance (0.655/0.648), ranking second in Video Quality (0.799/0.809), and first/second in Video



Figure 6: **Qualitative comparisons** on an ultra-long 139,210-frame source video. Target style: pixel art. When the content reappears after long temporal gaps, our approach preserves a consistent style and stable structures throughout, whereas competing approach SANA-Streaming changes the style drastically, and JoyAI-Video-Edit loses the style midway before drifting into a different one.

Table 2: **Ablation study** regarding the effectiveness of our proposed SISM, RSCR, and EA-SDM.

Configuration	IVEBench-Style (Short/Long)				CAM		IVEBench-Cycle	
	Total Score ↑	Video Quality ↑	Instruction Compliance ↑	Video Fidelity ↑	LPIPS ↓	DINO ↑	LPIPS ↓	DINO ↑
Ours	**0.758/0.756**	**0.799/0.809**	**0.655**/0.648	**0.821/0.811**	**0.2830**	**0.8051**	**0.0430**	**0.9786**
Ours w/o SISM	0.747/0.748	0.786/0.791	0.648/0.646	0.806/0.807	0.3977	0.7237	0.0769	0.9587
Ours w/o RSCR	0.751/0.753	0.796/0.800	0.649/**0.650**	0.809/0.810	0.3299	0.7815	0.0628	0.9690
Ours w/o EA-SDM	0.750/0.750	0.795/0.798	0.647/0.644	0.807/0.808	0.3416	0.7718	0.0563	0.9727

Fidelity (0.821/0.811). LiveEdit, EchoStyle, and StreamV2V preserve source content but apply weaker styles; SANA-Streaming and JoyAI-Video-Edit apply stronger styles, with JoyAI-Video-Edit closest overall (Total Score: 0.739/0.744). EverStyle thus improves Total Score while retaining strong visual quality and source fidelity. Fig. 3 shows the same balance: EverStyle expresses the target style while preserving main subjects, scene layouts, and fine structures across frames; other methods retain more source appearance or change structure more under stronger stylization. Users prefer EverStyle in 78% of pairwise comparisons (Appx. B.4), further supporting this balance.

Long-horizon consistency. As Tab. 1 (right) exhibits, EverStyle achieves leading results on both protocols. Compared with strongest competing scores, LPIPS/DINO improve from 0.0467/0.9741 to 0.0430/0.9786 on IVEBench-Cycle and from 0.4224/0.7060 to 0.2830/0.8051 on CAM, where gains are larger (33% lower LPIPS). Since IVEBench-Cycle measures accurate content recurrence whereas CAM evaluates natural scene revisitation under viewpoint changes, these results show consistent gains across two complementary settings. Qualitative comparisons also exhibit the advantages of our method. As shown in Fig. 4, StreamV2V drops rock-wall parts in later cycles, LiveEdit adds wall artifacts, SANA-Streaming changes colors of wall and clothing, and JoyAI-Video-Edit changes the right-hand subject's shoulder patterns, while EverStyle preserves appearance and structure across revisits. On CAM (Fig. 5), SANA-Streaming shifts the pixel-art palette and JoyAI-Video-Edit recolors the flowerbed across revisits, while StreamV2V and LiveEdit largely reproduce the source with weaker stylizations. On the hour-long, 139,210-frame stream (Fig. 6), JoyAI-Video-Edit gradually loses the target style, while SANA-Streaming gives the same street very different colors across distant revisits. EverStyle maintains a stable stylized appearance throughout, showing substantially stronger long-horizon consistency while preserving competitive stylization quality.

4.3 Ablation Studies

Component analysis. Tab. 2 validates the effectiveness of all three components. Removing SISM causes the largest degradation on both long-horizon benchmarks, with CAM performance dropping from 0.2830/0.8051 to 0.3977/0.7237, and IVEBench-Cycle from 0.0430/0.9786 to 0.0769/0.9587, demonstrating the importance of explicitly preserving historical stylization states. RSCR and EA-SDM provide complementary long-range supervision. Removing RSCR leads to a larger degradation on IVEBench-Cycle, where identical source content repeatedly reappears, while removing EA-SDM affects CAM more strongly, where scene revisits involve viewpoint variations and maintaining a stable global stylization becomes more important. Combining these three components consistently yields the strongest long-horizon consistency while preserving the best overall stylization quality.

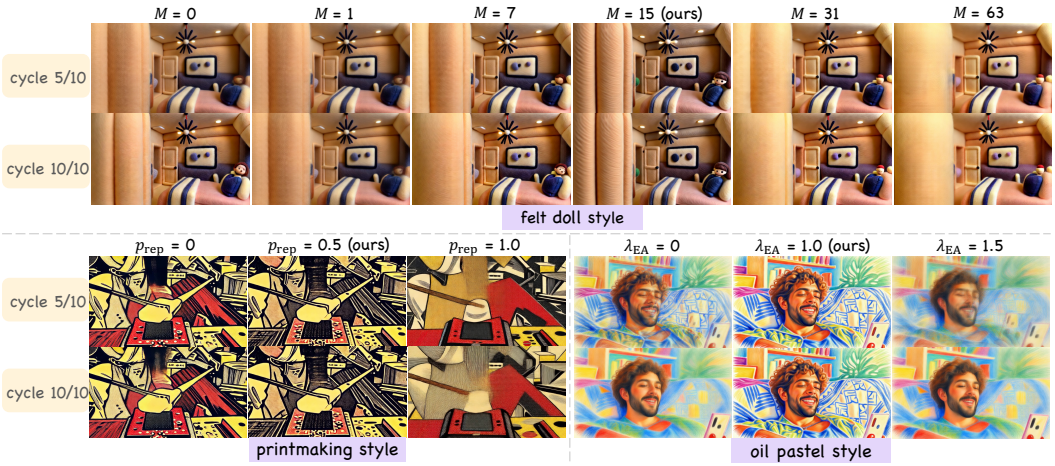

Figure 7: **Qualitative analysis of hyperparameters.** From top to bottom, we compare the results of various settings of SISM capacity M, RSCR activation probability p_{rep} and EA-SDM strength λ_{EA}. Each comparison exhibits corresponding outputs from the 5th and 10th rollout cycles. The selected default settings provide a favorable balance between long-horizon consistency and visual quality.

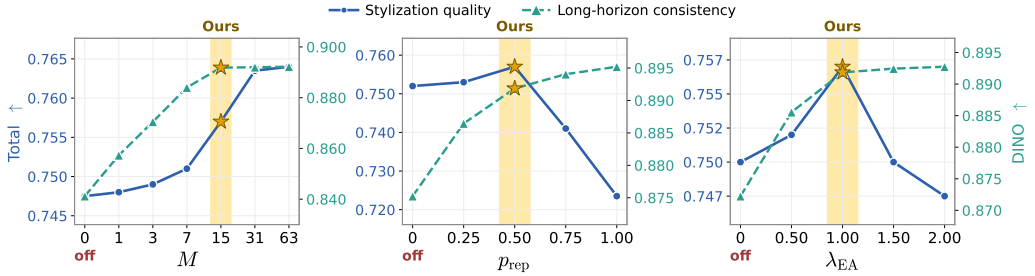

Figure 8: **Hyperparameter sensitivity analysis** for SISM capacity M, RSCR probability p_{rep}, and EA-SDM strength λ_{EA}, each varied while others stay at their default. Solid blue: generation quality (IVEBench Total Score over Short/Long); dashed green: long-horizon consistency (DINO over CAM and IVEBench-Cycle). Stars mark our final settings; "off" denotes disabling the component.

Hyperparameter sensitivity. We further study the effects of SISM capacity M, RSCR probability p_{rep}, and EA-SDM strength λ_{EA} in Figs. 7 and 8. Increasing M improves long-horizon consistency up to approximately $M = 15$, after which the gains largely saturate while the costs of storage and retrieval continue to increase. The qualitative results exhibit the same trend that insufficient memory fails to preserve recurring structures, whereas substantially larger banks can provide little additional consistency and may soften local details. For RSCR and EA-SDM, a moderate regularization gives the best balance. Disabling either objective leads to a visible appearance drift across distant cycles, whereas overly strong replay or anchoring suppresses fine details and reduces generation quality. Quantitatively, consistency generally improves as p_{rep} and λ_{EA} increase, but IVEBench quality peaks at the intermediate settings $p_{\text{rep}} = 0.5$ and $\lambda_{\text{EA}} = 1$. Moderate regularization thus gives a favorable balance: too little fails to prevent long-term drift, while too much constrains the model excessively and degrades stylization quality. We therefore use $M = 15$, $p_{\text{rep}} = 0.5$, and $\lambda_{\text{EA}} = 1$ for EverStyle.

Due to page limitations, more methodological details and experimental results are presented in **Appx./Supp.**

5 Conclusion

In this paper, we presented **EverStyle**, a streaming video stylization framework for maintaining consistent appearance over long causal rollouts. EverStyle addresses two key challenges of long-horizon inconsistency: the loss of historical stylization states once content leaves the active context, and the gradual drift of the global stylization over time. SISM and RSCR preserve and reinforce previously established stylizations when source content reappears, while EA-SDM constrains later generations against the style established earlier in the stream. Experiments demonstrate consistent improvements under both exact content recurrence and natural scene revisitation, while retaining competitive stylization quality. EverStyle further maintains stable stylizations on an hour-long, 139,210-frame rollout without resetting its history, suggesting that explicitly preserving historical stylization states and constraining long-range drift provides an effective basis for open-ended streaming video stylization.

REFERENCES

aigc apps. Videox-fun: A video generation pipeline for diffusion transformer, 2026. URL https://github.com/aigc-apps/VideoX-Fun.

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, Chenyang Zhang, Tao Xiang, Fanny Yang, Serge Belongie, and Tian Xie. Onestory: Coherent multi-shot video generation with adaptive memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16173–16184, June 2026.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Scaling instruction-based video editing with a high-quality synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pp. 707–723. Springer, 2022.

Peiliang Cai, Evelyn Zhang, Jiacheng Liu, Hao Lin, Ruiqi Zhang, Weile Mo, Yue Ma, Shikang Zheng, Jiehang Huang, Dongrui Liu, et al. Focused forcing: Content-aware per-frame kv selection for efficient autoregressive video diffusion. *arXiv preprint arXiv:2605.18346*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

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

Feng Chen, Bohan Zhuang, and Qi Wu. Streaming video diffusion: Online video editing with diffusion models. In *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–9. IEEE, 2025.

Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. Ivebench: Modern benchmark suite for instruction-guided video editing assessment. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026.

Michal Geyer, Omer Bar Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, volume 2024, pp. 1608–1620, 2024.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577, June 2025.

Nisha Huang, Yuxin Zhang, and Weiming Dong. Style-a-video: Agile diffusion for arbitrary text-based video style transfer. *IEEE Signal Processing Letters*, 31:1494–1498, 2024.

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.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, XiTai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-kv: Hybrid kv cache compression for efficient autoregressive video diffusion models, 2026. URL https://arxiv.org/abs/2605.09681.

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.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M. Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6507–6516, June 2024.

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

Dohun Lee, Chun-Hao Paul Huang, Xuelin Chen, Jong Chul Ye, Duygu Ceylan, and Hyeonho Jeong. Memory-v2v: Memory-augmented video-to-video diffusion for consistent multi-turn editing, 2026. URL https://arxiv.org/abs/2601.16296.

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

Huaqiu Li, Jiahao Wang, Sijia Cai, Hualian Sheng, Bing Deng, Jieping Ye, and Wenhan Luo. Echostyle: Unlocking high-fidelity video stylization with reverse data synthesis. *arXiv preprint arXiv:2606.25465*, 2026b.

Mengtian Li, Jinshu Chen, Songtao Zhao, Wanquan Feng, Pengqi Tu, and Qian He. Dreamstyle: A unified framework for video stylization. *arXiv preprint arXiv:2601.02785*, 2026c.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699, October 2025.

Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8207–8216, June 2024.

Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking backward: Streaming video-to-video translation with feature banks. In *International Conference on Learning Representations*, volume 2025, pp. 46425–46445, 2025.

Sen Liang, Fengbin Guan, Youliang Zhang, Xin Li, and Zhibo Chen. Cot-edit: Let cot guide instruction video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37960–37970, June 2026.

Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stabilizing shape consistency in video-to-video editing. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4):5467–5482, 2026a. doi: 10.1109/TCSVT.2025.3639307.

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*, volume 2026, pp. 91177–91196, 2026b.

Hengye Lyu, Zisu Li, Yue Hong, Yueting Weng, Jiaxin Shi, Hanwang Zhang, and Chen Liang. Dit as real-time rerenderer: Streaming video stylization with autoregressive diffusion transformer. In *European Conference on Computer Vision*, pp. 256–273. Springer, 2026.

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. Tethercache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026.

Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8089–8099, 2024.

POPtravel. Walking in SINGAPORE—chinatown to marina bay and gardens—4k 60fps (UHD). Video, Wikimedia Commons, 2019. URL https://commons.wikimedia.org/wiki/File:Walking_in_SINGAPORE_-_Chinatown_to_Marina_Bay_and_Gardens_-_4K_60fps_(UHD).webm. Published December 3, 2019. CC BY 3.0. Original video: https://www.youtube.com/watch?v=aUJl46bEWYo. Accessed September 18, 2026.

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 (ICCV)*, pp. 15932–15942, October 2023.

Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Quanjian Song, Mingbao Lin, Wengyi Zhan, Shuicheng Yan, Liujuan Cao, and Rongrong Ji. Univst: A unified framework for training-free localized video style transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, Siyu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7364–7373, June 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.

Xinyu Wang, Chongbo Zhao, Fangneng Zhan, and Yue Ma. Liveedit: Towards real-time diffusion-based streaming video editing. *arXiv preprint arXiv:2606.26740*, 2026.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7589–7599. IEEE, 2023.

Yicheng Xiao, Wenxun Dai, Xinran Qin, Lin Song, Maoquan Zhang, Hang Xu, Yukang Chen, Yitong Li, Guohui Zhang, Yuan Zhang, Xuying Zhang, Tommy Zhang, Jianlong Yuan, Peihao Li, Shuai Lu, Siming Fu, Chuyang Zhao, Xin Han, Jie Huang, Wenbo Li, Guoqing Ma, Wei Huang, Xiaojuan Qi, Haoyang Huang, and Nan Duan. Joyai-video-edit: Real-time open-ended video editing with autoregressive diffusion. *arXiv preprint arXiv:2608.03974*, 2026.

Zhening Xing, Gereon Fox, Yanhong Zeng, Xingang Pan, Mohamed Elgharib, Christian Theobalt, and Kai Chen. Live2diff: Live stream translation via uni-directional attention in video diffusion models. *arXiv preprint arXiv:2407.08701*, 2024.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8703–8712. IEEE, 2024.

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

Shuai Yang, Bingjie Gao, Ziwei Liu, Jiaqi Wang, Dahua Lin, and Tong Wu. Permavid: Consistent video generation across edits via disentangled context memory, 2026a. URL https://arxiv.org/abs/2606.16449.

Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yue Ma, Yan Huang, Min Xu, and Qiang Wu. Videocof: Unified video editing with temporal reasoner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37940–37949, 2026b.

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2630–2640, June 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

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

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, SA Conference Papers '25, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400721373. doi: 10.1145/3757377.3763833. URL https://doi.org/10.1145/3757377.3763833.

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. Memlearner: Learning to query context memory for video world models. In *European Conference on Computer Vision (ECCV)*, 2026.

Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. Veggie: Instructional editing and reasoning video concepts with grounded generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15147–15158, October 2025b.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zhenghao Zhang, Zuozhuo Dai, Long Qin, and Weizhi Wang. Effived: Efficient video editing via text-instruction diffusion models. *arXiv preprint arXiv:2403.11568*, 2024.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026a.

Yuyang Zhao, Yicheng Pan, Qiyuan He, Jincheng Yu, Junsong Chen, Tian Ye, Haozhe Liu, Enze Xie, and Song Han. Sana-streaming: Real-time streaming video editing with hybrid diffusion transformer. *arXiv preprint arXiv:2605.30409*, 2026b.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-cong Chen. Videomemory: Toward consistent video generation via memory integration. *arXiv preprint arXiv:2601.03655*, 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.

APPENDIX

This appendix provides more methodological details and experimental results for **EverStyle**. Sec. A presents the detailed design of SISM, including memory organization, verification-based writing, diversity-aware retention, confidence-calibrated retrieval, and correction-only fusion. Secs. B.1 and B.2 then provide more qualitative comparisons on stylization quality and long-horizon consistency, respectively. Sec. B.3 further examines the sensitivity to the main hyperparameters, while Sec. B.4 reports the protocol and results of our user study. Finally, Sec. C provides the AI use statement.

A DETAILED DESIGN OF SISM

Sec. 3.2 presents the high-level design of SISM. Here, we specify its memory operations in detail. As illustrated in Fig. 2 (b), SISM augments each self-attention layer with a fixed-budget memory of source-indexed stylization residuals. Each observation is first held as a pending candidate and committed only after a later observation verifies both source correspondence and stylization agreement. When the bank is full, entries are retained to favor diverse historical source content. During generation, SISM retrieves a historical residual only under a reliable source match and injects it through a separate correction branch, leaving the original sink-and-recent attention path intact. We detail these four operations below: verification-based writing, diversity-aware retention, confidence-calibrated retrieval, and correction-only fusion.

Per-layer memory organization. Recall from Eq. 3 that each source token i in a historical frame f is represented by a source address $\mathbf{a}_i^f$ and a stylization residual $\Delta\mathbf{v}_i^f$. For a frame with S tokens, we collect them as

$$\mathbf{A}^f = [\mathbf{a}_1^f, \dots, \mathbf{a}_S^f], \qquad \Delta\mathbf{V}^f = [\Delta\mathbf{v}_1^f, \dots, \Delta\mathbf{v}_S^f]. \tag{7}$$

A verified memory entry additionally stores token-wise reliability $\mathbf{r}^f = [r_1^f, \dots, r_S^f]$. The committed memory bank at each self-attention layer is therefore

$$\mathcal{B} = \{(\mathbf{A}_n^{\text{mem}}, \Delta\mathbf{V}_n^{\text{mem}}, \mathbf{r}_n^{\text{mem}})\}_{n=1}^{m}, \qquad m \le M, \tag{8}$$

together with one pending candidate. Thus, the total SISM state at each layer is bounded by M committed historical frames plus one candidate, independent of stream length.

Verification-based memory writing. Not every generated observation is suitable as a long-term reference, since a transient stylization error could otherwise be preserved and repeatedly reused. We therefore delay commitment: a new observation first enters the pending slot and is written into the bank only after a later observation confirms its stylization on overlapping source content.

Let f_{pend} denote the pending frame and f_{later} a later stylized frame. For each source token i in f_{pend}, we find its closest source token in f_{later}:

$$j_i^\star = \arg\max_j \mathbf{a}_i^{f_{\text{pend}}\top} \mathbf{a}_j^{f_{\text{later}}}, \qquad s_i^{\text{src}} = \mathbf{a}_i^{f_{\text{pend}}\top} \mathbf{a}_{j_i^\star}^{f_{\text{later}}}, \qquad \widehat{\Delta\mathbf{v}}_i^{f_{\text{later}}} = \Delta\mathbf{v}_{j_i^\star}^{f_{\text{later}}}. \tag{9}$$

Verification requires both source overlap and stylization agreement. For the former, we use a conservative lower-tail statistic of $\{s_i^{\text{src}}\}_{i=1}^{S}$, preventing a small coincident region from validating an otherwise unrelated frame. For the latter, we compare the aligned stylization residuals in direction and magnitude, yielding bounded agreement scores s_i^{dir} and s_i^{mag}.

When both criteria are satisfied, the pending observation is verified. We average the aligned residuals and assign each stored token a reliability score:

$$\Delta\mathbf{v}_i^{\text{mem}} = \frac{\Delta\mathbf{v}_i^{f_{\text{pend}}} + \widehat{\Delta\mathbf{v}}_i^{f_{\text{later}}}}{2}, \qquad r_i^{\text{mem}} = s_i^{\text{src}} \sqrt{s_i^{\text{dir}} s_i^{\text{mag}}}. \tag{10}$$

Averaging suppresses observation-specific generation noise, while r_i^{mem} captures both the quality of the source correspondence and the reproducibility of its stylization. The verified observation then becomes a candidate for retention. If verification fails, f_{later} replaces the current pending frame and is evaluated against subsequent observations instead.

Diversity-aware retention. Reliability alone does not determine which states deserve the limited memory capacity. Consecutive frames often contain highly similar source content, so retaining every

verified observation would quickly fill the bank with redundant entries. We instead favor coverage of diverse historical content.

If the bank contains fewer than M entries, a verified candidate is committed directly. Otherwise, let b^{cand} denote the candidate formed above. Among the existing entries and the new candidate, we retain

$$\mathcal{B}^\star = \arg \min_{\substack{\mathcal{S}\subseteq\mathcal{B}\cup\{b^{\text{cand}}\} \\ |\mathcal{S}|=M}} \max_{u\neq v\in\mathcal{S}} \text{Sim}_{\text{src}}(u, v), \tag{11}$$

where $\text{Sim}_{\text{src}}(u, v)$ measures frame-level source similarity from the stored addresses. Equivalently, among the $M + 1$ available entries, we remove the one whose removal yields the least redundant remaining bank. The new candidate is therefore retained only when it improves source coverage. Once committed, an entry remains unchanged until eviction, preventing later rollout drift from rewriting the historical reference.

Confidence-calibrated retrieval. Memory should affect generation only when the current source has a trustworthy historical correspondence. Since the sink and recent context are already available to the original attention path, SISM retrieves only from committed entries outside the active context. We flatten their stored addresses into memory tokens indexed by μ. For a current source token with address $\mathbf{a}$, we first retrieve

$$\mu^\star = \arg\max_{\mu} \mathbf{a}^\top \mathbf{a}^{\text{mem}}_\mu, \qquad \widetilde{\Delta\mathbf{v}} = \Delta\mathbf{v}^{\text{mem}}_{\mu^\star}, \qquad \widetilde{r} = r^{\text{mem}}_{\mu^\star}. \tag{12}$$

Nearest-neighbor retrieval alone is insufficient, since it always returns a match even when no genuine correspondence exists. Let $\rho^{(1)}$ and $\rho^{(2)}$ denote the highest and second-highest source similarities. We therefore calibrate each retrieved correction using its absolute match quality, matching margin, stored reliability, and relative correction magnitude:

$$\alpha = \mathcal{C}\Big(\rho^{(1)}, \rho^{(1)} - \rho^{(2)}, \widetilde{r}, e\Big), \qquad e = \log \frac{\text{RMS}(\widetilde{\Delta\mathbf{v}}) + \epsilon}{\text{RMS}(\mathbf{v}^{\text{src}}) + \epsilon}. \tag{13}$$

These four signals respectively indicate whether a strong match exists, whether it is unambiguous, whether the stored stylization was reliable when written, and whether the proposed correction has an appropriate scale. The lightweight calibrator $\mathcal{C}$ combines them into a confidence weight α, suppressing uncertain or excessive retrievals.

Correction-only fusion. A retrieved residual records *how* matched content was stylized historically, but its old spatial configuration should not determine *where* that stylization acts in the current frame. As shown in Fig. 2(b), we therefore preserve the original sink-and-recent attention as the base path and inject memory only through a separate correction branch aligned by the current source.

Let $\widetilde{\Delta\mathbf{V}}$ and $\boldsymbol{\alpha}$ collect the retrieved residuals and confidence weights for the current tokens. We compute

$$\begin{aligned} \mathbf{H}^{\text{base}} &= \text{Attn}\big(\mathbf{Q}^{\text{base}}, \mathbf{K}^{\text{base}}, \mathbf{V}^{\text{base}}\big), \\ \mathbf{H}^{\text{mem}} &= \text{Attn}\Big(\mathbf{Q}^{\text{base}}, \mathbf{K}^{\text{src}}, \boldsymbol{\alpha} \odot \widetilde{\Delta\mathbf{V}}\Big), \\ \mathbf{H} &= \mathbf{H}^{\text{base}} + \beta\mathbf{H}^{\text{mem}}, \qquad \beta = 0.5. \end{aligned} \tag{14}$$

Here, $\mathbf{Q}^{\text{base}}$, $\mathbf{K}^{\text{base}}$, and $\mathbf{V}^{\text{base}}$ are the queries, keys, and values of the original sink-and-recent attention, while $\mathbf{K}^{\text{src}}$ contains the current source keys. Using the current source keys aligns the historical correction with the present source geometry rather than its historical layout. When no reliable correspondence exists, confidence calibration suppresses the retrieved residual, causing the memory branch to vanish and recovering the original sink-and-recent generation path.

B Additional Experimental Results

B.1 Additional Stylization Quality Results

We provide additional qualitative comparisons on the Short and Long splits of IVEBench-Style under the evaluation protocol in Sec. 4.2. As shown in Fig. 9, LiveEdit, EchoStyle, and StreamV2V preserve source content well but tend to apply weaker stylization, whereas SANA-Streaming and

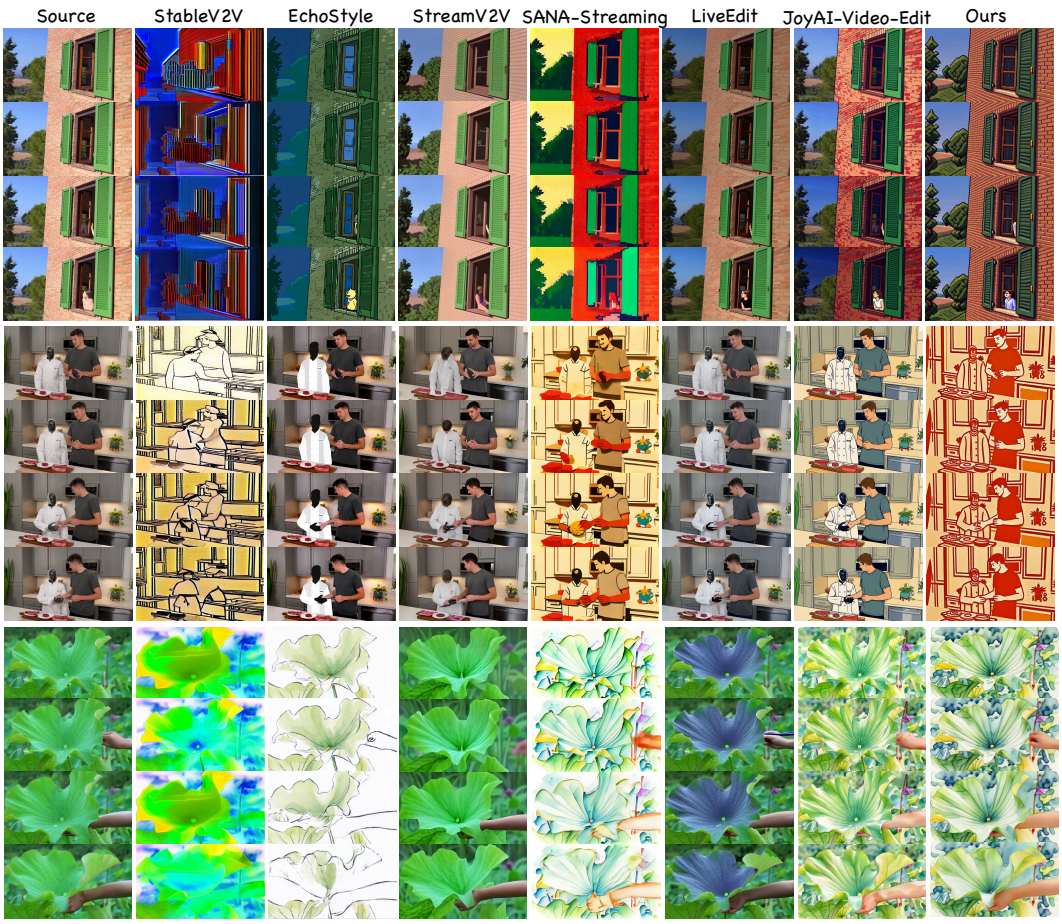

Figure 9: **Additional qualitative comparisons** on IVEBench-Style. Each group shows four representative frames from one test video. Target styles from top to bottom are pixel art, paper cutting, and watercolor. EverStyle applies the requested style clearly while preserving subjects, scene layout, and fine structural details across frames.

JoyAI-Video-Edit produce stronger style transformations with larger structural variations in some cases. In comparison, EverStyle expresses the target style clearly while better preserving subjects, scene layout, and fine structural details across the sampled frames. These results are consistent with the quantitative findings in Tab. 1, further demonstrating a favorable balance between style adherence and source fidelity on both IVEBench splits.

B.2 Additional Long-Horizon Stylization Consistency Results

We provide additional qualitative comparisons on IVEBench-Cycle under the exact content recurrence protocol in Sec. 4.2. We compare frames at matching within-cycle positions from the first, middle, and final cycles, such that each comparison corresponds to identical source content separated by long temporal gaps. As shown in Fig. 10, competing methods exhibit both appearance drift and structural inconsistency across cycles. In Case 2, JoyAI-Video-Edit changes the spatial relation between the person and the wall, while SANA-Streaming shows noticeable changes in both structure and color. In comparison, EverStyle preserves the established stylization and source structure across the sampled cycles. These results further demonstrate its ability to maintain consistent stylization under long-range content recurrence.

B.3 Additional Hyperparameter Ablations

We complement the quantitative sensitivity analysis in Sec. 4.3 with qualitative comparisons of SISM capacity M, RSCR activation probability p_{rep}, and EA-SDM strength λ_{EA}. For each study, we

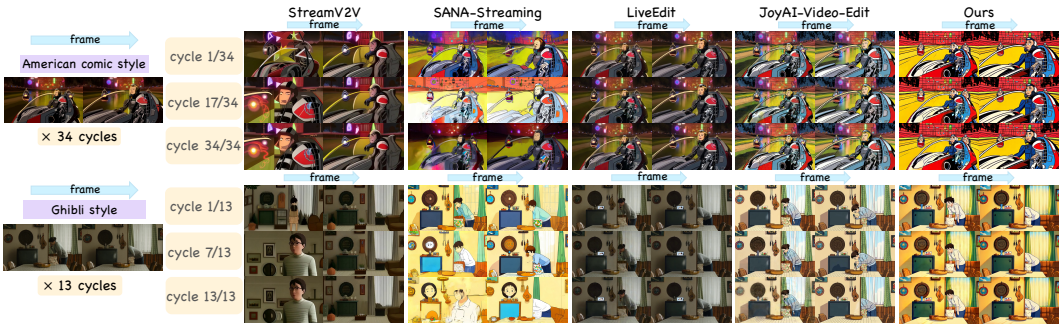

Figure 10: **Additional qualitative comparisons** on IVEBench-Cycle (source clip looped over 6,000 frames). For each case, rows show the first, 50%-position, and final cycles, with two frames at matching within-cycle positions. Our method preserves both the requested style and source structure throughout.

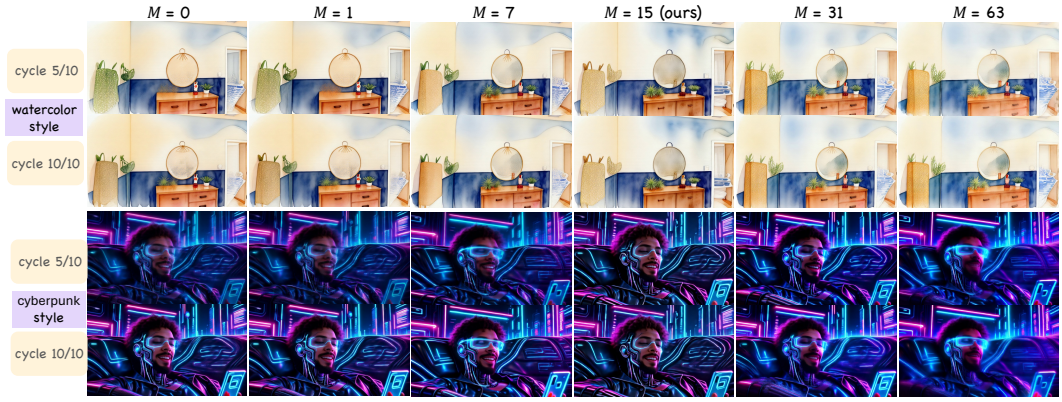

Figure 11: **Additional qualitative ablation** of the SISM capacity $M \in \{0, 1, 7, 15, 31, 63\}$. Corresponding frames are sampled from early and late portions of the same rollout. Increasing the capacity improves long-horizon consistency, with limited visible gains beyond the selected setting $M = 15$.

vary one hyperparameter while keeping the other two at their default values, using the same source video, style prompt, and sampled frame positions.

SISM capacity. As shown in Fig. 11, small memory capacities ($M = 0$, 1, and 7) lead to inconsistent stylization of recurring content. Increasing the capacity to $M = 15$ substantially improves consistency across distant recurrences, whereas larger capacities ($M = 31$ and 63) yield only limited additional gains. We therefore use $M = 15$, which provides sufficient historical coverage without unnecessary storage and retrieval overhead.

RSCR activation probability. As shown in Fig. 12, disabling RSCR ($p_{\text{rep}} = 0$) results in noticeable appearance changes across distant recurrences. Setting $p_{\text{rep}} = 0.5$ improves long-horizon consistency while preserving fine visual details, whereas applying replay to every training instance ($p_{\text{rep}} = 1$) weakens local texture and reduces visual quality. We therefore use $p_{\text{rep}} = 0.5$ to balance recurrence consistency and generation quality.

EA-SDM strength. As shown in Fig. 13, disabling EA-SDM ($\lambda_{\text{EA}} = 0$) leads to gradual changes in color and texture over the rollout. Setting $\lambda_{\text{EA}} = 1$ stabilizes the global stylization while retaining fine details, whereas a stronger constraint ($\lambda_{\text{EA}} = 1.5$) suppresses local details and softens object boundaries. We therefore use $\lambda_{\text{EA}} = 1$ to balance global style stability and visual quality.



Figure 12: **Additional qualitative ablation** of the RSCR activation probability $p_{\text{rep}} \in \{0, 0.5, 1.0\}$. Insufficient replay leads to inconsistent stylization across the rollout, whereas excessive replay reduces visual quality. We use $p_{\text{rep}} = 0.5$.

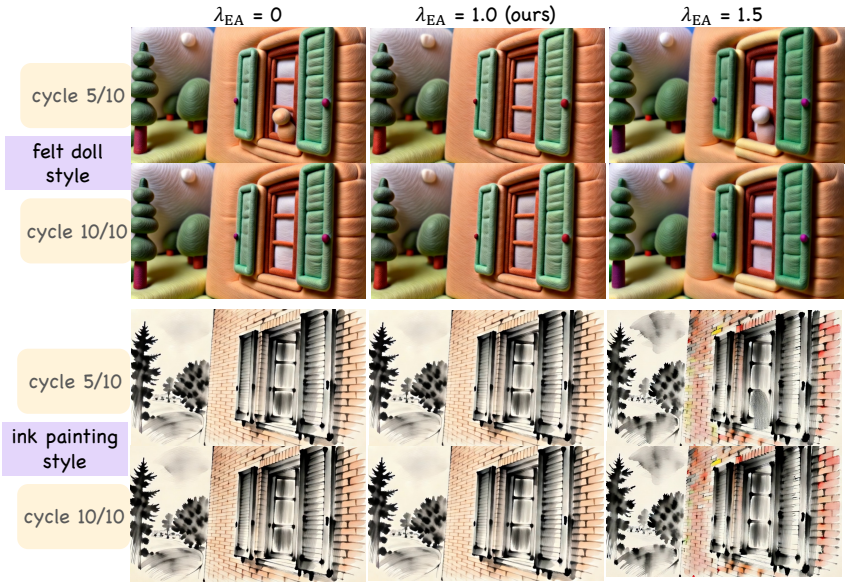

Figure 13: **Additional qualitative ablation** of the EA-SDM strength $\lambda_{\text{EA}} \in \{0, 1.0, 1.5\}$. Moderate anchoring stabilizes the global stylization state, while excessive strength suppresses fine details. We use $\lambda_{\text{EA}} = 1.0$.

B.4 User Study

To complement the automatic evaluation, we conduct a pairwise user study on perceptual stylization quality. The study includes 10 video cases and 7 methods, yielding $10 \times \binom{7}{2} = 210$ distinct method-pair comparisons. Under a balanced assignment, each participant evaluates 21 comparisons, covering every method pair once. Each question presents the source video together with corresponding outputs from two methods, and participants select "A is better", "Tie", or "B is better". We random-

Table 3: **Pairwise user-study results.** #Comp. denotes the number of comparisons involving each method.

Method	#Comp.	Win	Tie	Loss
StableV2V	120	23	42	55
EchoStyle	120	39	35	46
StreamV2V	120	11	41	68
SANA-Streaming	120	71	16	33
LiveEdit	120	8	46	66
JoyAI-Video-Edit	120	63	26	31
Ours	120	**94**	16	10

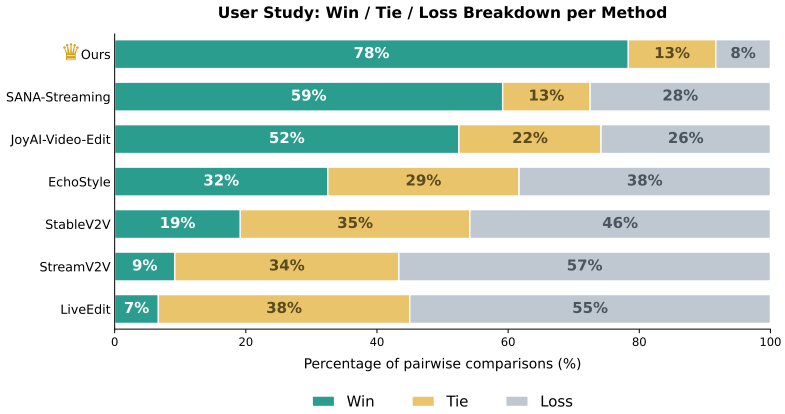

Figure 14: **Win/tie/loss proportions** for each approach, sorted by win rate. EverStyle achieves a 78.3% win rate.

ize both the left/right placement of the two methods and the question order to reduce positional and ordering biases.

As shown in Tab. 3 and Fig. 14, EverStyle receives 94 wins in 120 comparisons, corresponding to a 78.3% win rate. This is higher than SANA-Streaming (59.2%) and JoyAI-Video-Edit (52.5%), the two strongest competing methods in the study. The perceptual evaluation is consistent with our quantitative and qualitative results, providing additional evidence that EverStyle achieves a favorable balance of stylization quality and content preservation.

C AI Use Statement

AI tools were used only for language editing, including minor improvements to grammar, phrasing, and clarity. All technical content, including the ideas, methodology, experiments, and analyses, was developed by the authors, who take full responsibility for the manuscript.