WHEN THE FUTURE REMEMBERS: PAST RECONSTRUCTION IN AUTOREGRESSIVE LONG VIDEO GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: Eviction Forcing enables multi-minute long video generation while preserving long-range temporal consistency and high visual quality.

ABSTRACT

As autoregressive video diffusion models become the new paradigm for long-form video generation, a fundamental issue remains: these models tend to *forget* earlier content once it falls outside the bounded KV cache, resulting in identity drift and scene resets that compromise long-term temporal consistency. We propose **Eviction Forcing**, a simple yet effective training strategy that teaches the model to recover the past by conditioning on its own future. After evicting a prefix of frames from a standard base teacher rollout, a lightweight LoRA adapter is trained to reconstruct the evicted frames at their original RoPE indices, using only the remaining frames that are later on the temporal axis as context. This approach teaches the model to remember the past beyond the immediate attention window, thereby significantly improving video consistency across long durations. The model is further optimized via periodic supervision from the base teacher model under full-context rollouts, ensuring its stable performance during standard rollouts. Extensive experiments demonstrate that Eviction Forcing achieves competitive overall VBench scores and consistently improves long-range temporal consistency over state-of-the-art autoregressive baselines.

1 INTRODUCTION

Long-form video generation requires models to preserve subject identity, scene layout, motion continuity, and visual style over extended temporal horizons. Recent video diffusion models achieve strong visual quality on short clips by denoising all frames jointly with bidirectional attention, but this full-sequence formulation introduces substantial memory and computational overhead. Autoregressive (AR) diffusion models offer an effective alternative: each frame is generated by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

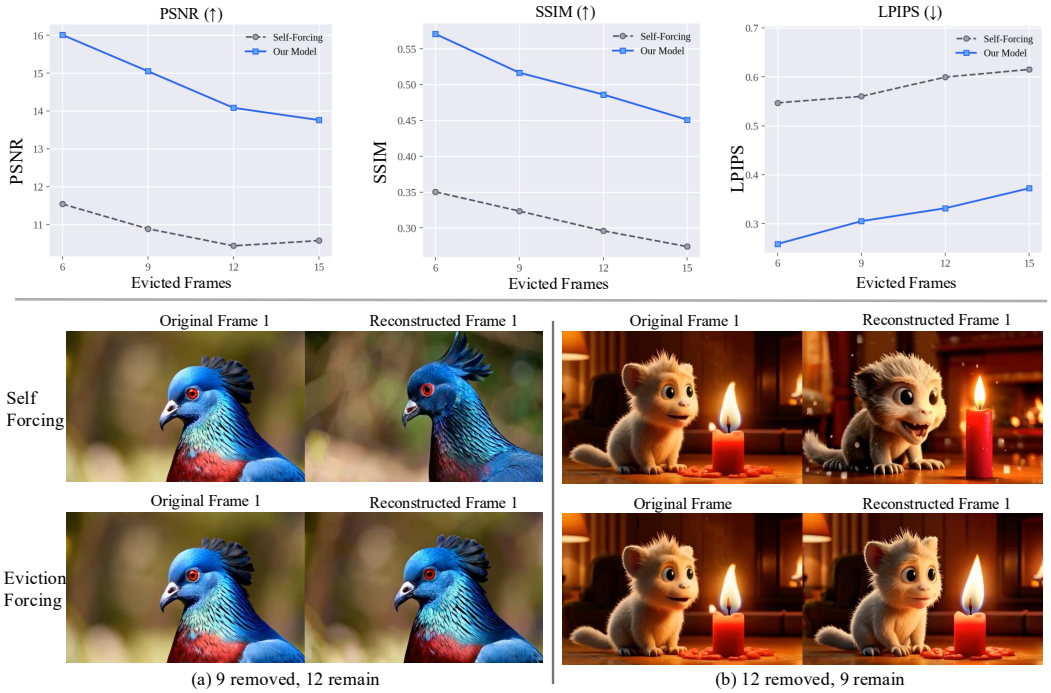

Figure 2: **Motivation** - past reconstruction from future context. **Top:** Quantitative comparison of the reconstruction. Eviction Forcing consistently outperforms Self Forcing across all three metrics. **Bottom:** Qualitative comparison of the reconstruction. Self Forcing exhibits noticeable identity and background drift, whereas Eviction Forcing faithfully preserves both subject identity and background across both eviction depths.

attending only to previously generated ones, and key-value (KV) cache enables efficient reuse of past context. These properties make AR diffusion well suited for streaming and long-form video generation.

Despite these advantages, long-horizon AR generation remains fundamentally limited by the size of the KV cache. Existing approaches either reduce train-test mismatch, as in Self Forcing (Huang et al., 2025), or engineer the inference context through sinks, re-caching, rolling windows, or positional encoding. For example, Self-Forcing++ (Cui et al., 2025) re-injects noise for teacher correction, Rolling Forcing (Liu et al., 2025) uses a joint denoising rolling window, and Infinity-RoPE (Yesiltepe et al., 2025) remaps temporal positions. These mechanisms control which context is retained, processed, or indexed, but do not teach the weights to recover information once its source frames leave the cache. We instead treat long-term identity, scene, and motion preservation as a learned capacity of the generator itself: surviving frames were generated from the evicted prefix, still carry traces of it, and can teach the model to recover the past in weight space.

This weight-space perspective leads to a different question: rather than asking how to keep the past in memory, *can the model learn to reconstruct it from its own future?* The premise follows naturally from AR generation that later frames are produced while attending to earlier ones and therefore implicitly encode the traces of their identity, motion, and layout. To test whether these traces are recoverable, we perform an **Eviction-Reconstruction** diagnostic. Given a standard 21-frame rollout, we evict a prefix from the KV cache and reconstruct it at its original RoPE indices using only the remaining later frames, whose indices are left unchanged. As shown in Fig. 2 (bottom), Eviction Forcing outperforms the base model on the reconstruction task, substantially improving the preservation of subject identity and background structure. Quantitative results in Fig. 2 (top) further demonstrate that across all eviction depths in $\{6, 9, 12, 15\}$, Eviction Forcing more faithfully reconstructs the perceptual content of past frames, achieving higher PSNR and SSIM (Wang et al., 2004) and lower LPIPS (Zhang et al., 2018) compared to the baseline. These results confirm that the surviving contains usable evidence of the evicted past, motivating it as a direct training signal.

In this work, we propose **Eviction Forcing**, a simple yet effective Eviction-Reconstruction training strategy that teaches the model to recover the past while conditioning solely on its own future. Given a standard rollout from a frozen short-horizon base teacher, we evict a prefix of frames from the KV cache while preserving the RoPE positions of the remaining frames. We then train a lightweight LoRA adapter to regenerate the evicted frames at their original RoPE positions, using only the non-evicted frames as context, which serve as future to the reconstructed past. To maintain standard full-context generation performance, we further introduce a **Behavior Preservation** module that periodically supervises full-length LoRA rollouts against the base teacher. Crucially, Eviction Forcing is a weight-space training method, not a cache, sink, or positional-encoding mechanism. At inference, we adopt Infinity-RoPE (Yesiltepe et al., 2025), which keeps temporal RoPE indices within the trained horizon. While Infinity-RoPE supports long-horizon generation, it cannot leverage information lost to KV cache eviction. Eviction Forcing closes this gap by learning to recover past content from its own future, thereby preserving long-range consistency beyond what RoPE remapping alone can achieve.

Our contributions are summarized as follows:

- We identify a previously overlooked asymmetry in autoregressive video generation: future frames implicitly preserve information about past frames, providing a supervisory signal for maintaining long-horizon consistency.
- We propose **Eviction Forcing**, a weight-space training strategy that enables the model itself to recover past information from its own future, improving consistency in long-form video generation.
- We conduct extensive qualitative and quantitative experiments on VBench-Long, demonstrating that Eviction Forcing outperforms state-of-the-art autoregressive video generation methods on overall metrics.

2 Related Work

Bidirectional Video Generation. Recent progress in video generation has been driven primarily by diffusion models (Lipman et al., 2022; Song et al., 2020). From early pixel-space formulations to latent-space approaches, modern diffusion models are commonly built on DiT backbones (Peebles & Xie, 2023). Large-scale video diffusion models, such as HunyuanVideo (Kong et al., 2024), Movie Gen (Polyak et al., 2024), CogVideoX (Yang et al., 2024), and Wan (Wan et al., 2025) have advanced video generation to remarkable visual quality. However, their bidirectional attention mechanisms require each frame to attend to every other frame, incurring heavy computational and memory overhead as well as high latency. These limitations prevent real-time and long-horizon deployment, which have motivated a shift toward causal, autoregressive formulations for video generation.

Autoregressive Video Diffusion. Autoregressive video diffusion decomposes video generation into a chain of conditional denoising steps over past frames, enabling streaming and long-form video generation (Yin et al., 2025; Huang et al., 2025; Cui et al., 2025; Li et al., 2026a; Liu et al., 2025; Cui et al., 2026; Zhao et al., 2026; Chen et al., 2026; Xiang et al., 2026; Lu et al., 2025; Po et al., 2025; Yesiltepe et al., 2026; Bian et al., 2026). CausVid (Yin et al., 2025) distills a bidirectional DiT into a causal few-step student model via DMD, but its train-test mismatch induces exposure bias beyond the teacher's training horizon. Self-Forcing (Huang et al., 2025) addresses this by conditioning on its own generated frames during training, yet the generation quality still degrades beyond the short training horizon. Self-Forcing++ (Cui et al., 2025) extends generation to longer horizons by re-injecting noise into self-generated long rollouts and applying DMD on short segments, enabling teacher correction without long-horizon training. LongLive (Yang et al., 2025) introduces streaming long tuning for extended rollouts and KV-recache for prompt transitions. Rolling Forcing (Liu et al., 2025) employs a joint denoising window that simultaneously denoises multiple frames with increasing noise levels, reducing error accumulation over long horizons. To improve motion dynamics in long rollouts, Reward Forcing (Lu et al., 2025) introduces an EMA-sink for ultra-long generation and Rewarded DMD for enhanced motion dynamics. A parallel line of work targets long-horizon inference through RoPE modifications (Yesiltepe et al., 2025; Cui et al., 2026; Li et al., 2026b). Infinity-RoPE (Yesiltepe et al., 2025) reformulates 3D-RoPE as a relativistic moving frame to overcome the 1024-frame indexing limit. FLEX (Li et al., 2026b) further bridges the gap between short training and long inference by interpolating under-trained low-frequency dimensions while extrapolating high-frequency components.

3 BACKGROUND

Autoregressive Video Diffusion Models. Our method builds on Wan2.1-T2V-1.3B (Wan et al., 2025) distilled from Self Forcing (Huang et al., 2025). AR video diffusion generates videos sequentially: each new latent frame or chunk is denoised while conditioning only on previously generated ones. This gives the causal factorization:

$$p(x_{1:N}) = \prod_{i=1}^{N} p(x^i \mid x_{<i}), \tag{1}$$

where $x_{<i}$ denotes the previously generated frames. Each conditional distribution is modeled by a diffusion- or flow-based denoising process. Under the rectified-flow formulation, a noisy latent is formed by interpolating between a clean latent x_0^i and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ as $x_t^i = (1-t)x_0^i + t\epsilon$, where $t \in [0, 1]$. The reverse process is parameterized by a neural velocity field v_θ conditioned on the previously generated context, with $\frac{dx_t^i}{dt} = v_\theta(x_t^i, t, x_{<i})$. At inference, we integrate from noise at $t = 1$ to a clean latent at $t = 0$ using a small number of Euler steps.

3D Rotary Position Embedding (3D-RoPE). Wan2.1 (Wan et al., 2025) employs 3D-RoPE to encode the temporal and spatial coordinates of query and key tokens to distinguish them across positions in time, height, and width. For the temporal axis, a query $\mathbf{q}$ at position m and a key $\mathbf{k}$ at position n are rotated according to their temporal indices before attention. The rotation is given by $\langle \mathbf{q}'_m, \mathbf{k}'_n \rangle = \langle \mathbf{q}_m, \mathbf{R}(n-m)\mathbf{k}_n \rangle$, where $\mathbf{R}(\cdot)$ denotes the RoPE rotation. Therefore, the attention score depends solely on the relative positional distance $n - m$.

4 METHODOLOGY

4.1 EVICTION-RECONSTRUCTION

Autoregressive video generation uses the KV cache as a temporal memory. During long rollouts, when newly generated frames are appended to the cache, older frames are removed due to bounded KV cache. Once early frames are evicted, the model no longer has direct access to their visual information including subject, background, and motion. Eviction-Reconstruction trains the model under this exact failure mode. We intentionally remove a prefix from the KV cache and ask the model to reconstruct the evicted frames using only the remaining later frames.

As illustrated in Fig. 3(a), Eviction-Reconstruction consists of three steps. First, a frozen base teacher produces a standard autoregressive rollout. Second, we evict a prefix from the resulting KV cache while preserving the temporal RoPE positions of the remaining frames. Third, a LoRA adapter reconstructs the evicted prefix from the post-eviction cache. The base model is frozen throughout training, and only the LoRA adapter is optimized.

Base Teacher Rollout. We begin by disabling the LoRA adapter and running the frozen short-horizon base model in its standard autoregressive mode. The model generates a full-context rollout:

$$X = (x_1, x_2, \ldots, x_N)$$

where each x_i denotes a clean latent frame or latent block. After each block is generated, its keys and values are written into the KV cache and reused for subsequent generation. This matches the autoregressive rollout observed at inference.

This teacher rollout serves two purposes. First, it produces clean target latents, which are used as supervision targets for LoRA-conditioned reconstruction. This keeps the LoRA adapter aligned with the distribution produced by the model it will modify. Second, it populates the KV cache with entries that the model would observe at inference, ensuring that the cache state seen by the LoRA adapter during training is distributionally consistent with inference.

Prefix Eviction. After the teacher rollout, we split the KV cache into two parts:

$$X_{\text{evict}} = (x_1, \ldots, x_n), \qquad X_{\text{keep}} = (x_{n+1}, \ldots, x_N)$$

We evict X_{evict} from the KV cache and retain only the cache entries corresponding to X_{keep}. The eviction depth n is uniformly sampled from $\{6, 9, 12\}$ at each training step. This sampling strategy

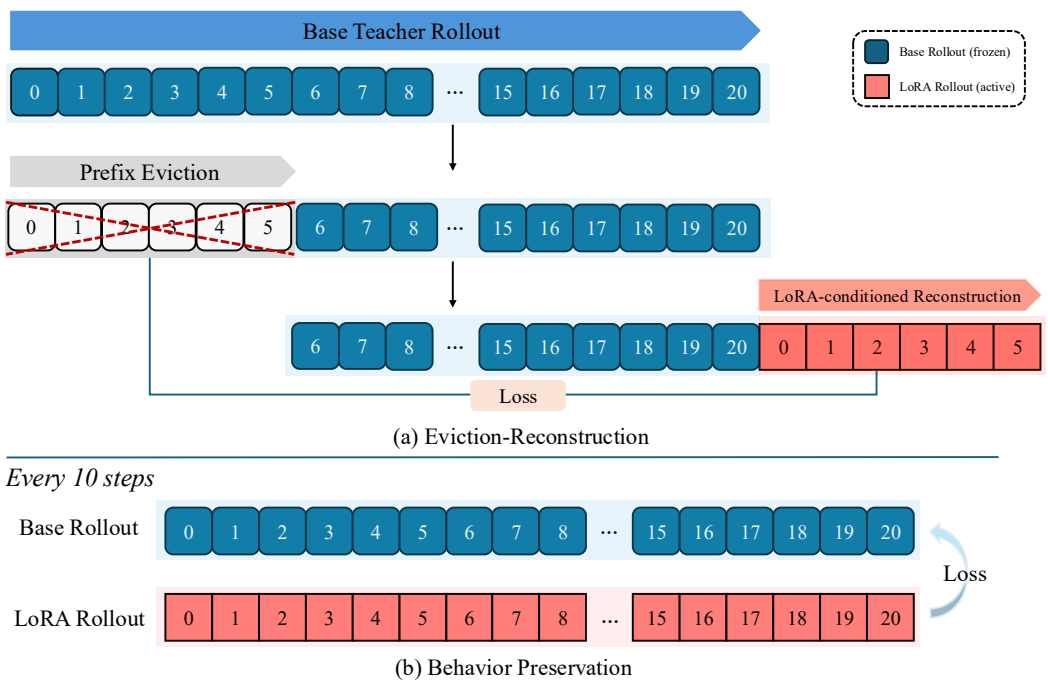

Figure 3: **Overview of Eviction Forcing. (a) Eviction-Reconstruction** evicts a prefix of n frames from a base model rollout, e.g., 6 frames from a 21-frame rollout, and trains a LoRA adapter to reconstruct the evicted frames at their original RoPE positions using only the remaining later frames. **(b) Behavior Preservation** periodically supervises full-context LoRA rollouts against the frozen base teacher to preserve standard autoregressive generation behavior.

ensures that shallow, medium, and deep eviction depths can be handled simultaneously by the same LoRA adapter. X_{keep} are shifted leftward to fill the empty cache positions. Despite occupying smaller cache positions, their RoPE indices are kept unchanged $[n, \ldots, N-1]$ under 0-indexing.

Maintaining the original temporal coordinates for non-evicted frames is essential for the subsequent reconstruction task. Re-indexing the remaining frames according to their new cache positions would break the past-from-future conditioning structure that our objective aims to explore: when querying past positions $[0, 1, \ldots, n-1]$, the LoRA adapter must attend to keys whose temporal coordinates are later than those of the frames being reconstructed. Preserving the original RoPE indices ensures that this temporal asymmetry remains visible to the attention mechanism.

LoRA-conditioned Reconstruction. We then enable the LoRA adapter to reconstruct the evicted prefix. The adapter is asked to regenerate $\hat{X}_{\text{evict}}$ at the original RoPE positions of X_{evict}, while conditioning only on the post-eviction cache from X_{keep}, which corresponds to future positions relative to the reconstructed frames. This reverses the usual autoregressive direction. We ask the adapter to predict the evicted past from later frames that were generated from them.

This task is meaningful because the later frames are not independent of the evicted prefix. During the teacher rollout, frames in X_{keep} were generated while attending to X_{evict}. As a result, they carry implicit information of the removed frames, including subject identity, scene composition, and motion dynamics. Thus, the adapter learns to extract past information from its own future.

We supervise reconstruction by adopting two complementary objectives: a flow-matching MSE loss that aligns the LoRA-adapted model's velocity prediction with the rectified-flow target defined by the teacher's clean output, and an LPIPS loss (Zhang et al., 2018) that preserves fine perceptual details to which MSE is insensitive.

We define the flow-matching loss over a target sequence X and context C as:

$$\mathcal{L}_{\text{FM}}(X; C) = \mathbb{E}_{x \in X,\, t,\, \epsilon}\left[\left\| v_\theta\big((1-t)x + t\epsilon,\, t;\, C\big) - (\epsilon - x)\right\|_2^2\right] \tag{2}$$

where x is a clean teacher latent, $\epsilon \sim \mathcal{N}(0, I)$, and C denotes the cache used as conditioning context. The target velocity $\epsilon - x$ follows the rectified-flow formulation.

The reconstruction objective is then defined as:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{FM}}(X_{\text{evict}}; C_{\text{keep}}) + \lambda_{\text{lpips}}^{\text{recon}} \mathcal{L}_{\text{LPIPS}} \left(\hat{X}_{\text{evict}}, X_{\text{evict}} \right) \tag{3}$$

where C_{keep} is the post-eviction cache containing only the remaining future frames, and $\lambda_{\text{lpips}}^{\text{recon}}$ balances the LPIPS term against the flow-matching loss.

4.2 Behavior Preservation

While Eviction-Reconstruction trains the LoRA adapter exclusively under the post-eviction scheme of recovering the past by querying the future, it does not explicitly enforce stable generation quality under standard full-context rollouts. Therefore, LoRA may over-specialize to the reconstruction and alter the base model's standard generation behavior. To address this limitation, we introduce a **Behavior Preservation** module that is invoked every T_p training steps.

At each preservation step, we reset the KV cache and seed it with a full autoregressive rollout produced entirely by the LoRA-augmented model, matching the horizon of the base's rollout. This rollout is initialized with the same noise z used by the teacher model, ensuring that any divergence between the LoRA output and the teacher target is attributed to the LoRA adapter and supervised against the teacher's clean latents.

Let X_{full} denote the teacher rollout and $\hat{X}_{\text{full}}$ denote the corresponding LoRA rollout. We define the preservation loss with the same two terms as reconstruction:

$$\mathcal{L}_{\text{pres}} = \lambda_p \left[\mathcal{L}_{\text{FM}}(X_{\text{full}}; \hat{C}_{\text{full}}) + \lambda_{\text{lpips}}^{\text{pres}} \mathcal{L}_{\text{LPIPS}} \left(\hat{X}_{\text{full}}, X_{\text{full}} \right) \right] \tag{4}$$

where $\hat{C}_{\text{full}}$ is the standard full-context autoregressive cache built from $\hat{X}_{\text{full}}$, λ_p controls the contribution of the preservation objective, and $\lambda_{\text{lpips}}^{\text{pres}}$ weights the LPIPS term.

Behavior Preservation acts as an anchor to the base model. It keeps the LoRA adapter compatible with the base model's standard generation behavior. This is important because Eviction-Reconstruction changes the usual attention pattern by allowing earlier RoPE positions to query later ones. Periodic full-context supervision counterbalances this shift and preserves the standard autoregressive behavior.

The final training objective is then defined as:

$$\mathcal{L}_{\text{total}}(t) = \mathcal{L}_{\text{recon}} + \mathbf{1}_{\{t \bmod T_p = 0\}} \mathcal{L}_{\text{pres}} \tag{5}$$

where t is the training step and $\mathbf{1}_{\{\, t \bmod T_p = 0 \,\}}$ activates the preservation loss every T_p steps. Thus, reconstruction is optimized at every step, while preservation is applied periodically to anchor the adapter under full-context rollouts.

5 Experiments

Implementation Details. We implement Eviction Forcing on top of Self Forcing (Huang et al., 2025), a causal generator distilled from Wan2.1-T2V-1.3B (Wan et al., 2025), which generates videos of 21 latent frames with resolution 832×480. We train the LoRA adapter with rank 64. All experiments use KV cache size 21 and block size 3. The eviction depth is uniformly sampled from $\{6, 9, 12\}$. The preservation loss weight λ_p, reconstruction LPIPS loss weight $\lambda_{\text{lpips}}^{\text{recon}}$, and preservation LPIPS loss weight $\lambda_{\text{lpips}}^{\text{pres}}$ are all set to 0.1. Training is conducted on a single NVIDIA A100 80GB GPU for 24 hours with a batch size of 1. We use AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 2.5×10^{-5}. At inference, we implement Eviction Forcing with Infinity-RoPE (Yesiltepe et al., 2025).

Baselines. We compare Eviction Forcing against several state-of-the-art autoregressive video generation models, including Self Forcing (Huang et al., 2025), Causal Forcing (Zhu et al., 2026), LongLive (Yang et al., 2025), Infinity-RoPE (Yesiltepe et al., 2025), Rolling Forcing (Liu et al., 2025), Deep Forcing (Yi et al., 2025), and Reward Forcing (Lu et al., 2025).

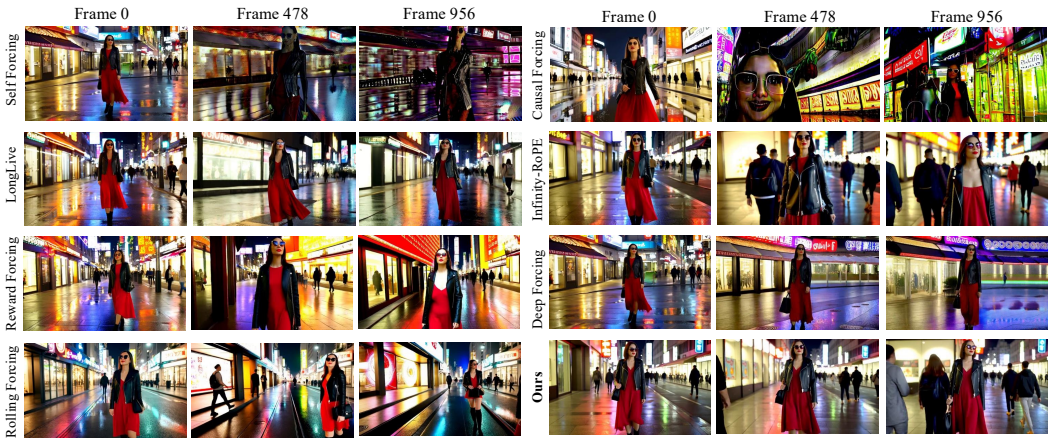

Figure 4: **Qualitative results on 60-second videos.** Baselines exhibit identity and scene drift, color shift, temporal inconsistencies, and degraded visual quality. Eviction Forcing strongly preserves subject identity and spatial layout while maintaining temporal consistency with high frame quality throughout the generation horizon.

Evaluation. We evaluate our method on VBench-Long (Huang et al., 2024). It measures multiple dimensions of video generation, including aesthetic quality, background consistency, dynamic degree, imaging quality, motion smoothness, subject consistency, and temporal flickering. We randomly sample prompts from MovieGenBench (Polyak et al., 2024) and generate over 100 videos across four durations: 5, 60, 120, and 240 seconds.

Training and Inference Cost. Eviction Forcing is a single lightweight adaptation: the 1.3B backbone is frozen and only a rank-64 LoRA (23.6M parameters, 1.82%) is optimized. Each step uses a gradient-free 21-frame teacher rollout and a gradient pass only through the 6, 9, or 12 evicted frames. Preservation is amortized to once every $T_p = 10$ steps. This incurs only a small training cost. At inference, the LoRA merges into the base weights with zero overhead, preserving Self Forcing's architecture, cache and block sizes, and denoising steps. Infinity-RoPE adds only negligible index remapping, leaving throughput unchanged at approximately 17 FPS on one H100.

5.1 Main Results

Qualitative Results. All video results are shared in the supplementary material. We provide 60-second results in Fig. 4 and 120-second results in Fig. 7, Appendix A.3. Across both horizons, Eviction Forcing better preserves subject identity, including facial structure, hairstyle, and clothing, while baselines exhibit subject swaps, appearance drift, or garment inconsistencies. Eviction Forcing also demonstrates strong temporal consistency, with subject motion and scene content evolving smoothly across frames. In contrast, several baselines exhibit abrupt discontinuities, background shifts, blurriness, artifacts, or over-saturation during long rollouts. See Appendix A.3 for further discussion.

Quantitative Results. We evaluate Eviction Forcing on 5s, 60s, 120s, and 240s generation. As shown in Table 1, Eviction Forcing remains competitive on 5s videos, ranking second in background consistency, subject consistency, and motion smoothness. This indicates that our model preserves short-horizon performance. On 60s videos, Eviction Forcing ranks first on all three metrics, substantially mitigating identity, scene, and motion drift when extrapolating beyond the training horizon. Table 2 further shows that Eviction Forcing achieves the second-best overall score at 120s and the best overall score at 240s. Although Reward Forcing leads at 120s due to an aggressive Dynamic Degree, Eviction Forcing offers a more balanced trade-off across temporal consistency, aesthetic quality, and motion richness. At 240s, Eviction Forcing achieves the best overall score and highest subject consistency, demonstrating its robustness under extreme rollout lengths.

Table 1: **Performance comparisons on 5s and 60s videos. AQ**: Aesthetic Quality, **BC**: Background Consistency, **DD**: Dynamic Degree, **IQ**: Imaging Quality, **MS**: Motion Smoothness, **SC**: Subject Consistency, **TF**: Temporal Flickering. **Bold**: best; underline: second best. Our model is competitive on 5s generation, and on 60s generation it ranks first on BC, SC, MS, and TF while achieving the second-best overall score.

Model	Results on 5s↑								Results on 60s↑							
	AQ	BC	DD	IQ	MS	SC	TF	Overall	AQ	BC	DD	IQ	MS	SC	TF	Overall
Baselines																
Self Forcing (Huang et al., 2025)	0.618	0.961	0.625	0.682	0.984	0.970	0.968	0.830	0.565	0.958	0.393	0.650	0.987	0.974	0.978	0.787
Causal Forcing (Zhu et al., 2026)	0.623	0.957	**0.792**	**0.710**	0.976	0.964	0.953	**0.853**	0.503	0.936	**0.847**	0.608	0.935	0.920	0.918	0.810
Rolling Forcing (Liu et al., 2025)	0.610	0.960	0.292	0.701	**0.988**	0.975	0.978	0.786	0.580	0.958	0.380	0.670	0.988	0.977	0.979	0.790
Deep Forcing (Yi et al., 2025)	0.617	0.961	0.625	0.682	0.984	0.970	0.968	0.830	0.597	0.957	0.402	**0.690**	0.987	0.979	0.978	0.799
Reward Forcing (Lu et al., 2025)	**0.628**	0.958	0.646	0.686	0.986	0.965	0.972	0.834	0.585	0.952	0.676	0.673	0.985	0.974	0.971	**0.831**
LongLive (Yang et al., 2025)	0.626	**0.969**	0.271	0.684	0.991	**0.979**	**0.983**	0.786	**0.606**	0.961	0.433	0.664	**0.991**	0.982	0.982	0.803
Infinity-RoPE (Yesiltepe et al., 2025)	0.607	0.963	0.542	0.669	0.987	0.974	0.977	0.817	0.607	0.959	0.577	0.638	0.988	0.979	0.979	0.818
Ours	0.612	0.965	0.556	0.662	0.988	0.976	0.978	0.820	0.604	**0.970**	0.545	0.646	**0.991**	**0.988**	**0.984**	0.818

Table 2: **Performance comparisons on 120s and 240s videos.** On 120s generations, our model leads on AQ, BC, and SC. On 240s generations, our model achieves the highest overall score, demonstrating robust long-horizon preservation without sacrificing motion richness.

Model	Results on 120s↑								Results on 240s↑							
	AQ	BC	DD	IQ	MS	SC	TF	Overall	AQ	BC	DD	IQ	MS	SC	TF	Overall
Baselines																
Self Forcing (Huang et al., 2025)	0.541	0.948	0.624	0.577	0.952	0.932	0.933	0.787	0.542	0.957	0.426	0.620	0.979	0.954	0.969	0.778
Causal Forcing (Zhu et al., 2026)	0.526	0.945	0.738	0.628	0.968	0.947	0.950	0.814	0.487	0.954	0.557	0.562	0.964	0.945	0.943	0.773
Rolling Forcing (Liu et al., 2025)	0.620	0.953	0.742	**0.688**	0.982	0.960	0.968	0.845	0.592	0.961	0.590	**0.711**	0.984	0.967	0.972	0.825
Deep Forcing (Yi et al., 2025)	0.621	0.953	0.713	0.660	0.979	0.961	0.965	0.836	0.637	0.955	0.650	0.653	0.974	0.962	0.958	0.827
Reward Forcing (Lu et al., 2025)	0.644	0.956	**0.954**	0.683	0.981	0.957	0.965	**0.877**	**0.639**	0.962	0.742	0.685	0.984	0.965	0.967	0.849
LongLive (Yang et al., 2025)	0.654	0.959	0.649	0.678	**0.987**	0.967	**0.976**	0.839	0.639	**0.968**	0.377	0.688	**0.988**	0.972	**0.978**	0.801
Infinity-RoPE (Yesiltepe et al., 2025)	0.640	0.958	0.847	0.669	0.982	0.966	0.968	0.862	0.626	0.959	0.633	0.672	0.984	0.968	0.971	0.832
Ours	**0.672**	**0.968**	0.793	0.686	0.983	**0.974**	0.975	0.864	0.604	0.967	**0.770**	0.691	0.983	**0.974**	0.969	**0.851**

5.2 Ablation Studies

Ablation on KV Cache Size. Our default setting runs a 21-frame rollout with eviction depths sampled from $\{6, 9, 12\}$. We additionally train two variants: KV cache size of 9 with eviction depths sampled from $\{3, 6\}$, and KV cache size of 15 with eviction depths sampled from $\{6, 9\}$. Eviction ratios are kept roughly comparable across three variants so that the LoRA adapter faces similar reconstruction difficulty across settings. As shown in Table 3, all three variants achieve comparable overall scores on 60s generation, indicating the scalability of our method that is invariant to the choice of context length. We do observe a small advantage of the default 21-frame setting at 120s generation. We attribute this to the richer temporal context enabled by the larger cache size, while noting that our method still remains effective under shorter context lengths.

Ablation on Eviction Depths. Our default training uniformly samples the eviction depth from $\{6, 9, 12\}$ per example. To evaluate the effect of this distributional training strategy, we compare against three single-depth eviction variants in which n is fixed to 6, 9, and 12 for each training. As shown in Table 3, the default setting outperforms all single-depth variants on both 60s and 120s generations. We also observe that no single eviction depth consistently dominates the others, indicating that fixed-depth training generalizes poorly across the broader generation distribution. In contrast, uniformly sampled training learns a depth-conditional reconstruction strategy, producing a robust adapter across a wide range of eviction depths encountered at inference.

Ablation on Base Teacher Supervision. Our default training uses the base model to supervise LoRA reconstruction. To evaluate the contribution of base supervision, we run two ablations that replace the base target with a LoRA-on rollout. The first is LoRA self-distillation, which removes the base teacher entirely and reconstructs LoRA's own outputs after prefix eviction. The second adds a weak teacher signal by applying periodic preservation every 10 steps. As shown in Table 3, removing the teacher entirely causes a consistent drop across nearly all metrics, demonstrating the benefit of pairing the teacher signal with the Eviction-Reconstruction objective. Periodic base anchoring yields modest gains, indicating that it is helpful but not a sufficient substitute for using the base teacher as the reconstruction target.

Table 3: **Ablation studies** on KV Cache Size, Eviction Depth, and Base Teacher Supervision.

Model	Results on 60s ↑								Results on 120s ↑							
	AQ	BC	DD	IQ	MS	SC	TF	Overall↑	AQ	BC	DD	IQ	MS	SC	TF	Overall↑
Ablation on KV Cache Size																
Cache size 9	0.604	0.966	0.549	0.642	0.989	0.982	0.981	0.816	0.655	0.967	0.720	0.640	0.985	0.974	0.974	0.845
Cache size 15	0.609	0.968	0.556	0.647	0.989	0.982	0.980	0.819	0.659	0.967	0.728	0.633	0.985	0.973	0.972	0.845
Ablation on Eviction Depth																
Eviction depth 6	0.596	0.967	0.528	0.625	0.989	0.982	0.980	0.809	0.603	0.965	0.721	0.646	0.982	0.973	0.972	0.838
Eviction depth 9	0.609	0.967	0.450	0.636	0.990	0.983	0.981	0.802	0.657	0.965	0.765	0.641	0.982	0.972	0.971	0.851
Eviction depth 12	0.602	0.969	0.487	0.623	0.989	0.983	0.981	0.805	0.661	0.965	0.721	0.637	0.982	0.972	0.972	0.844
Ablation on Base Teacher Supervision																
LoRA self-dist.	0.596	0.965	0.503	0.613	0.979	0.983	0.977	0.802	0.659	0.965	0.733	0.640	0.985	0.972	0.973	0.847
LoRA self dist. + pres.	0.602	0.968	0.528	0.617	0.979	0.983	0.981	0.808	0.661	0.967	0.694	0.642	0.984	0.973	0.973	0.842
Ours	0.604	0.970	0.545	0.646	0.991	0.988	0.984	0.818	0.672	0.968	0.793	0.686	0.983	0.974	0.975	0.864

Table 4: **Transferability and sensitivity on 60s generation. Left:** Eviction Forcing across three backbones. **Right:** one-at-a-time hyperparameter sweeps (**bold:** best).

Backbone	Variant	SC	BC	Overall
Self Forcing	Base	0.974	0.958	0.787
	+Eviction Forcing	**0.988**	**0.970**	**0.818**
CausVid	Base	0.977	0.946	0.774
	+Eviction Forcing	**0.985**	**0.963**	**0.804**
SkyReels-V2	Base	0.969	0.954	0.782
	+Eviction Forcing	**0.984**	**0.970**	**0.813**

Hyperparameter	Value	Overall
Rank	32	0.815
	64	0.818
	128	0.820
T_p	5	0.812
	10	0.818
	20	0.815
λ_{lpips}	0	0.808
	0.1	0.818
	0.5	0.816

Transferability. We apply the same training recipe, without backbone-specific tuning, to two additional backbones: DMD-distilled CausVid and diffusion-forcing SkyReels-V2-DF-1.3B (Chen et al., 2025), evulated on 60s generation. As shown in Table 4 (left), Eviction Forcing consistently improves subject consistency, background consistency, and overall score across all three backbones, suggesting that it addresses shared post-eviction information loss rather than a specific architecture or cache design. **Hyperparameter Sensitivity.** We vary LoRA rank, preservation interval T_p, and λ_{lpips} one at a time while keeping others at their defaults. Table 4 (right) shows that lower ranks modestly reduce model capacity, while doubling to 128 yields only marginal gains at twice the adapter size. Too-frequent preservation over-anchors to the base, whereas too-sparse preservation weakens the anchor. Removing LPIPS degrades consistency, while any nonzero weights are comparable. Overall, the untuned defaults remain at or near optimal, indicating that observed gains come from the training objective itself rather than hyperparameter tuning.

5.3 User Study

We conducted a user study with 20 participants evaluating 60s and 120s generated videos on a 5-point Likert scale across four aspects: subject and scene consistency, motion quality, visual quality and aesthetics, and overall quality. Table 5 reports mean scores for each duration. Eviction Forcing achieves the highest average scores at both durations, with leading consistency scores consistent with our quantitative results. See Appendix A.2 for further discussion.

Table 5: **User study. SC**: Subject & Scene Consistency, **MQ**: Motion Quality, **AQ**: Aesthetic Quality, **OVQ**: Overall Quality, **Avg.**: mean across metrics.

Model	60s↑					120s↑				
	SC	MQ	AQ	OVQ	Avg.	SC	MQ	AQ	OVQ	Avg.
Baselines										
Self Forcing (Huang et al., 2025)	1.10	1.16	1.06	1.10	1.11	1.03	1.17	1.04	1.06	1.08
Causal Forcing (Zhu et al., 2026)	1.04	1.11	1.05	1.05	1.06	1.09	1.20	1.06	1.10	1.11
Rolling Forcing (Liu et al., 2025)	2.21	2.60	2.03	2.13	2.24	2.33	2.68	2.13	2.78	2.48
Deep Forcing (Yi et al., 2025)	1.99	2.68	2.04	2.68	2.35	2.01	2.38	2.11	2.16	2.17
Reward Forcing (Lu et al., 2025)	2.40	2.85	2.25	2.67	2.54	2.54	2.81	2.34	2.67	2.59
LongLive (Yang et al., 2025)	3.65	**3.80**	**3.48**	3.51	3.61	3.58	3.65	**3.59**	3.32	3.54
Infinity-RoPE (Yesiltepe et al., 2025)	3.13	3.56	3.34	3.26	3.32	3.51	3.60	3.38	3.54	3.51
Ours	**4.16**	3.77	3.45	**3.60**	**3.75**	**3.94**	**3.85**	3.53	**3.59**	**3.73**

6 Conclusion

We propose Eviction Forcing, a lightweight framework for long-form autoregressive video generation under a bounded KV cache. By training a LoRA adapter to recover past context from its own future, Eviction Forcing preserves past information beyond the immediate attention window. This enables accurate reconstruction of evicted context and maintains temporal consistency over ultra-long rollouts. Overall, Eviction Forcing offers a simple weight-space framework for temporally coherent long-form video generation.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks with required disclosure. We used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. All AI-suggested edits were reviewed by authors to ensure that they did not alter the original content or claims of the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We use the publicly available MovieGenBench prompts under their license. This work does not involve human subjects, potentially harmful insights, potential conflicts of interest and sponsorship, discrimination or fairness concerns, privacy and security issues, legal compliance, or research integrity issues. All authors adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of the experimental results presented in this work, we provide details of the experimental setup and implementation needed for reproducing the reported results in the Implementation Details under the Experiments section. Additionally, we will release the source code upon publication.

REFERENCES

Yuxuan Bian, Zeyue Xue, Songchun Zhang, Shiyi Zhang, Weiyang Jin, Yaowei Li, Junhao Zhuang, Haoran Li, Jie Huang, Haoyang Huang, et al. Echo-infinity: Learning evolving memory for real-time infinite video generation. *arXiv preprint arXiv:2606.04527*, 2026.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

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

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. *arXiv preprint arXiv:2512.12080*, 2025.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Xunzhi Xiang, Zixuan Duan, Guiyu Zhang, Haiyu Zhang, Zhe Gao, Junta Wu, Shaofeng Zhang, Tengfei Wang, Qi Fan, and Chunchao Guo. Pathwise test-time correction for autoregressive long video generation. *arXiv preprint arXiv:2602.05871*, 2026.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

Hidir Yesiltepe, Jiazhen Hu, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, Hoda Eldardiry, and Pinar Yanardag. Videomla: Low-rank latent kv cache for minute-scale autoregressive video diffusion. *arXiv preprint arXiv:2605.30351*, 2026.

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

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

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.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed kv-memory for consistent long video generation. *arXiv preprint arXiv:2603.21366*, 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.

A Appendix

A.1 Videos and Website

We provide an extensive collection of generated visual results, including videos showcasing motivations, qualitative performance, ablation studies, and side-by-side comparisons against baseline models. Please check videos in the supplementary file `website.html`.

A.2 User Studies

We show the interface used for our user study in Fig 5 and Fig 6, each for 60 seconds and 120 seconds generations. For each prompt, we include videos generated by our model and all other baselines for fair comparison. Users were presented with the prompts and generated results and were asked to evaluate each video from four different aspects that are designed to be mutually distinct and jointly comprehensive: (1) overall video quality, measuring how well the video matches the prompt, how realistic the content looks, and its overall success as a complete generation; (2) subject and scene consistency, measuring how the main subject preserves its identity, and whether the background and overall scene structure remain stable throughout the duration; (3) motion quality, measuring whether motions appear realistic and continuous, and whether the video is free from jitter, flickering, or abrupt changes; and (4) overall visual quality and aesthetics, measuring sharpness, clarity, color balance, and overall aesthetic appeal of the video.

We recruited 20 participants, each rating videos generated from 5 different prompts per duration, with all eight methods (seven baselines and Eviction Forcing) per prompt. Ratings were collected on a 5-point Likert scale (1 = Very Bad, 5 = Very Good). As shown in Table 5, Eviction Forcing achieves the highest average score at both durations, surpassing the strongest baselines, LongLive Yang et al. (2025) and Infinity-RoPE Yesiltepe et al. (2025). Consistent with our quantitative findings, our model achieves the strongest performance in subject and scene consistency. We further observe that Eviction Forcing leads on overall video quality at both durations, suggesting the significance of temporal consistency for perceptually higher-quality generations.

A.3 More Discussion on Results

A.3.1 Discussion on 60 Seconds Generation

Eviction Forcing shows its primary advantage when the generation horizon substantially exceeds the training horizon. As shown in Table 1, at 60-second generation, Eviction Forcing leads on consistency-related metrics including background consistency, subject consistency, motion smoothness, and temporal flickering, while remaining comparable on the others. These four metrics are the ones that capture the failure modes that typically emerge under long-horizon rollouts, therefore serving as diagnostic indicators of long-horizon generation quality.

The advantage of our model is more pronounced in qualitative comparisons. As demonstrated in Fig. 4, we observe several distinct failure modes among the baselines at 60 seconds durations. Reward Forcing and Infinity-RoPE exhibit severe identity drift throughout the entire rollout, with facial features, clothing details, and body proportions gradually distorting over time. LongLive shows a similar failure mode, with facial details beginning to deform mid-rollout. Deep Forcing preserves subject identity well but suffers from substantial background distortion. Rolling Forcing struggles with overall video quality, including subject swaps, background collapse, and motion discontinuity. In comparison, our model maintains stable subject identity and a consistent background throughout the entire 60-second window without sacrificing motion smoothness.

These qualitative observations align directly with the quantitative results. The specific failure mode of each baseline corresponds to the metrics on which it underperforms, while the strong consistency preservation of Eviction Forcing is reflected in its leading scores across all four consistency-related metrics.

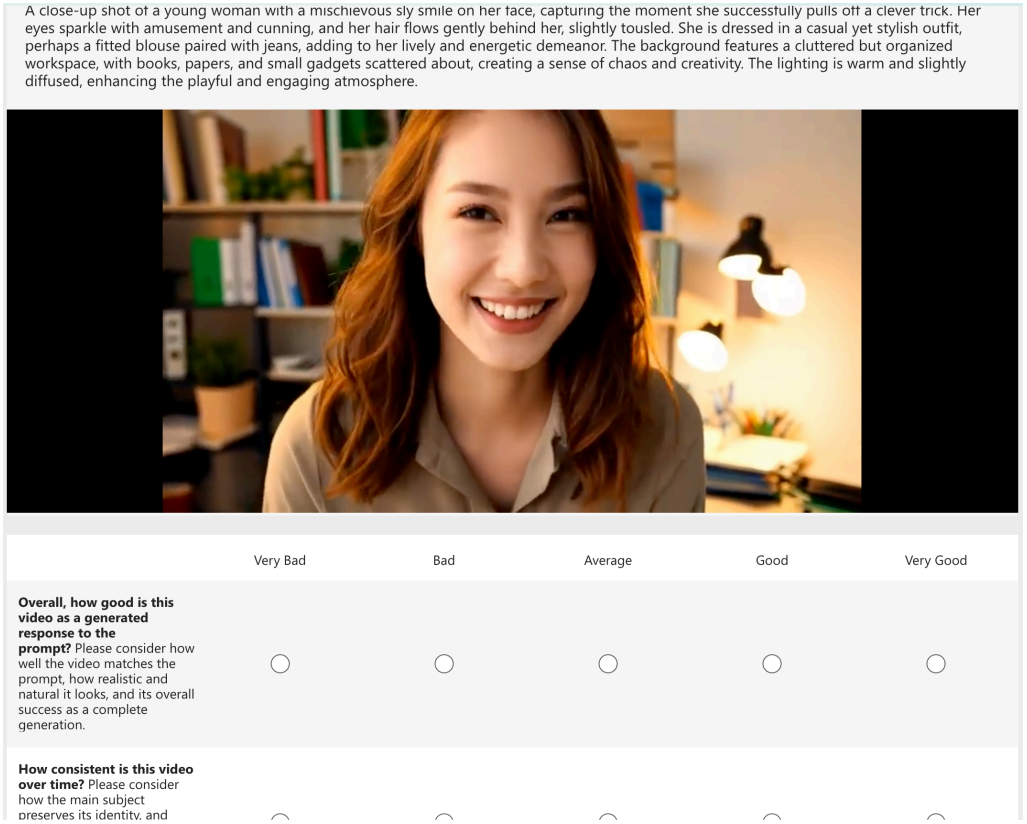

Figure 5: **User study.** User study interface for 60 seconds generation.

A.3.2 Discussion on 120 Seconds Generation

At 120-second generation, Eviction Forcing quantitatively sustains its performance and widens the gap over the baselines. Our method continues to lead on the consistency-related metrics observed at 60 seconds and additionally improves on aesthetic-oriented metrics including aesthetic quality and image quality.

The contrast between Eviction Forcing and the baselines becomes more evident qualitatively at this duration. As shown in Figure 7, our model preserves the overall scene layout and, more importantly, fine-grained subject details including facial features, hairstyle, and the floral pattern on the dress throughout the full 120-second horizon. Eviction Forcing achieves this level of detail preservation without sacrificing motion smoothness or dynamics, both of which remain comparable to the baselines. In contrast, baseline outputs at this duration exhibit additional failure modes beyond consistency collapse: Rolling Forcing shows visible structural artifacts that appear unnatural within the scene, while Deep Forcing exhibits oversaturated foreground colors that break the overall scene coherence. Other baselines suffer from the same subject deformation, background distortion, and temporal flickering observed in the 60 seconds rollout. Among the baselines, Infinity-RoPE is the strongest at this duration, preserving identity and overall layout reasonably well. However, the floral pattern on the dress drifts continuously throughout the rollout (as shown in Fig 8), indicating that even the strongest baseline struggles to maintain fine-grained visual details over long horizons.

A.3.3 Discussion on 240 Seconds Generation

At 240-second generation, our method achieves the best overall score, leading in subject consistency and dynamic degree while ranking second in background consistency, imaging quality, and temporal flickering. Our method achieves this without sacrificing motion dynamic for consistency, a trade-off that affects most approaches at this duration. This balance is particularly significant because 240-

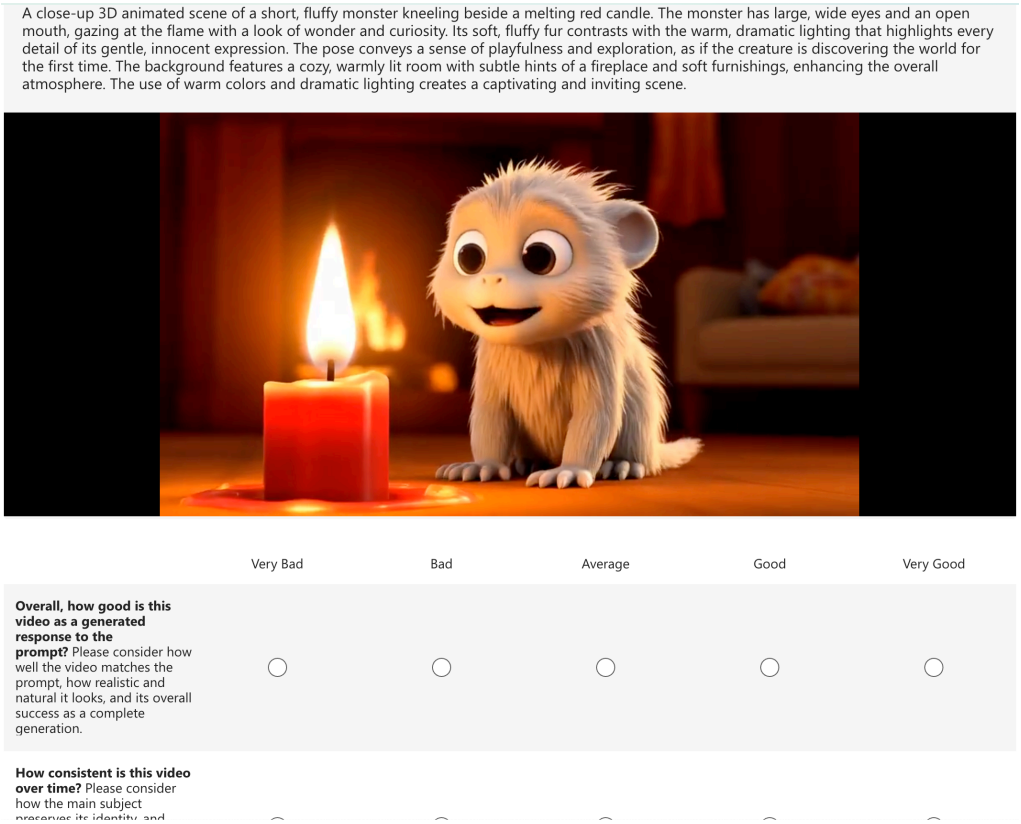

Figure 6: **User study.** User study interface for 120 seconds generation.

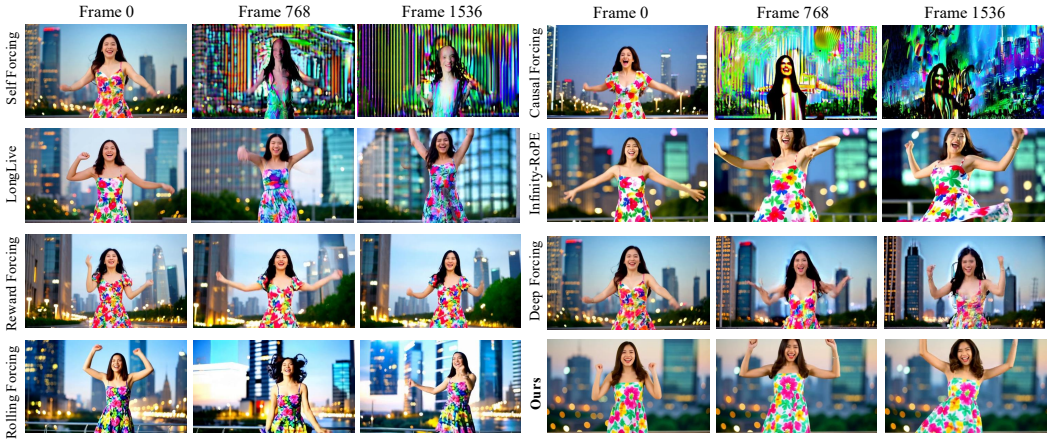

Figure 7: **Qualitative results on 120-second videos.** Baseline models exhibit subject drift, background distortions, temporal inconsistency, and degraded visual quality. Our model demonstrates strong subject preservation, consistent background structure, and temporal consistency with high frame quality throughout the horizon.

second rollouts extend 48 $\times$ beyond the training context, placing all methods in difficulty where consistency and dynamism typically trade off against one another.

The contrast between our model and other baselines is most visible in the 240 seconds comparisons in the supplementary material. Our model maintains coherent subject identity, scene composition, and motion smoothness across the entire four-minute rollout. Notably, our model again retains

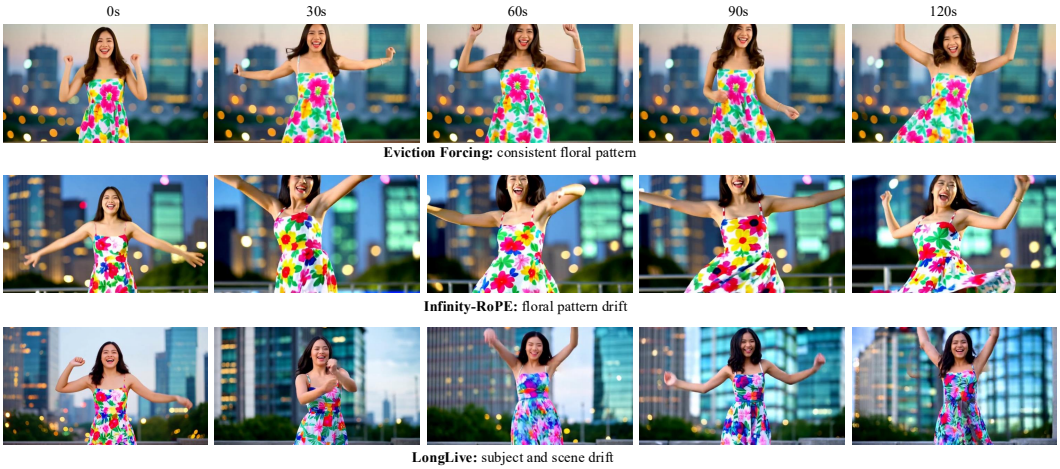

Figure 8: **Comparison on 120s generation.** Frames sampled uniformly at 0s, 30s, 60s, 90s, and 120s. Eviction Forcing preserves the subject's floral pattern, identity, and background throughout. Infinity-RoPE Yesiltepe et al. (2025) drifts in the floral pattern and LongLive Yang et al. (2025) loses both subject identity and scene composition over time.

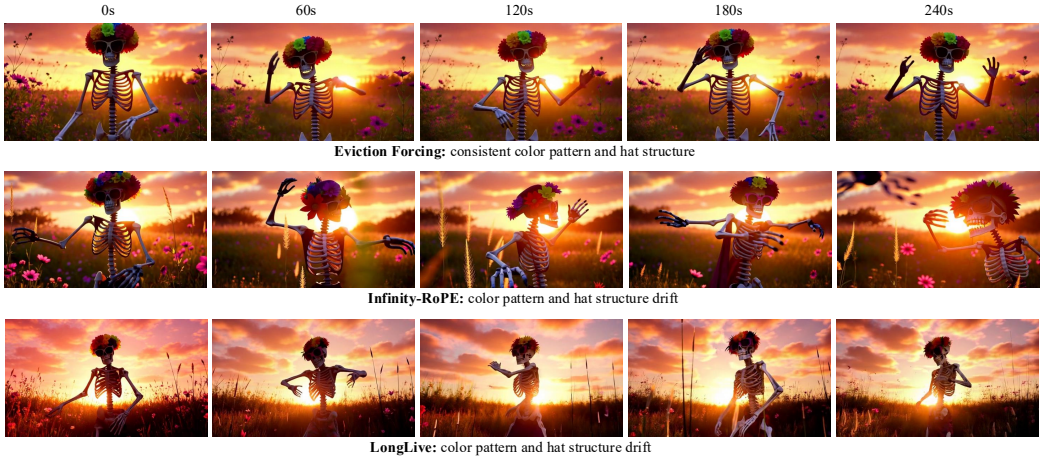

Figure 9: **Comparison on 240s generation.** Frames sampled uniformly at 0s, 60s, 120s, 180s, and 240s. Eviction Forcing preserves both the color pattern and hat structure throughout the rollout, while Infinity-RoPE Yesiltepe et al. (2025) and LongLive Yang et al. (2025) exhibit progressive color and structure drift over time.

fine-grained details: the flower hat composed of multiple colors remains its exact pattern and color composition throughout the four-minute rollout (as shown in Fig 9), whereas several baseline models show color shifting over time.