Relax Forcing: Relaxed KV-Memory for Consistent Long Video Generation

Zengqun Zhao[1], Yanzuo Lu[2], Ziquan Liu[1], Jifei Song[3], Jiankang Deng[2], Ioannis Patras[1]

[1]Queen Mary University of London, [2]Imperial College London, [3]Huawei R&D UK
https://zengqunzhao.github.io/Relax-Forcing

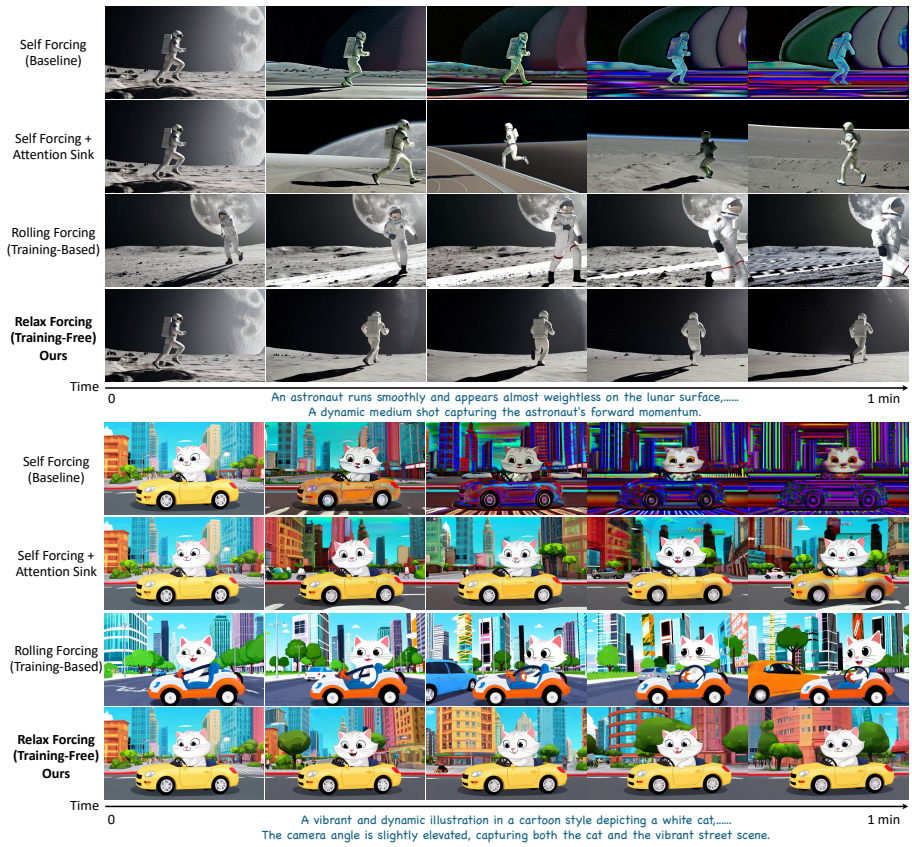

Fig. 1: Long-horizon autoregressive video generation over one minute. Baseline Self Forcing suffers from severe temporal drift and visual degradation as errors accumulate. Adding an attention sink stabilises early frames but limits motion diversity. Rolling-Forcing improves short-term coherence but still exhibits inconsistent patterns. In contrast, our Relax-Forcing dynamically selects informative history while suppressing redundancy, achieving stable identity preservation and sustained scene evolution.

Abstract. Autoregressive (AR) video diffusion has recently emerged as a promising paradigm for long video generation, enabling causal synthesis beyond the limits of bidirectional models. To address training–inference

mismatch, a series of self-forcing strategies have been proposed to improve rollout stability by conditioning the model on its own predictions during training. While these approaches substantially mitigate exposure bias, extending generation to minute-scale horizons remains challenging due to progressive temporal degradation. In this work, we show that this limitation is not primarily caused by insufficient memory, but by how temporal memory is utilised during inference. Through empirical analysis, we find that increasing memory does not consistently improve long-horizon generation, and that the temporal placement of historical context significantly influences motion dynamics while leaving visual quality largely unchanged. These findings suggest that temporal memory should not be treated as a homogeneous buffer. Motivated by this insight, we introduce Relax Forcing, a structured temporal memory mechanism for AR diffusion. Instead of attending to the dense generated history, Relax Forcing decomposes temporal context into three functional roles: Sink for global stability, Tail for short-term continuity, and dynamically selected History for structural motion guidance, and selectively incorporates only the most relevant past information. This design mitigates error accumulation during extrapolation while preserving motion evolution. Experiments on VBench-Long demonstrate that Relax Forcing improves motion dynamics and overall temporal consistency while reducing attention overhead. Our results suggest that structured temporal memory is essential for scalable long video generation, complementing existing forcing-based training strategies.

Keywords: Long video generation · AR diffusion models · KV memory

1 Introduction

Recent advances in video diffusion models have enabled high-fidelity synthesis of short videos [1,17,31,33]. To support interactive [26,32,38] and streaming [2, 12, 34] applications that require minute-scale generation, autoregressive (AR) diffusion models have emerged as a practical solution, generating frames causally with key–value (KV) caching mechanisms [3,13,21,29,37]. Rollout-based training strategies, such as Self Forcing [13] and its subsequent extensions [5, 6, 22, 23, 40, 46], mitigate exposure bias by allowing the model to condition on its own generated frames during training. During inference, these models extrapolate beyond the training horizon by autoregressively rolling out generation in a sliding window manner, where newly generated frames are appended to the context while older frames are discarded. This enables videos longer than the training sequence length to be synthesised.

While rollout-based training substantially improves robustness, extending AR diffusion to minute-scale synthesis remains fundamentally challenging, a problem commonly referred to as long-video extrapolation [43,44]. Even with reduced training–inference mismatch, autoregressive models must repeatedly condition on previously generated frames, allowing residual errors and temporal biases to accumulate over time. Consequently, long-horizon generation often exhibits either gradual drift or overly constrained motion dynamics. Recent work therefore shifts attention from training strategies to memory management during

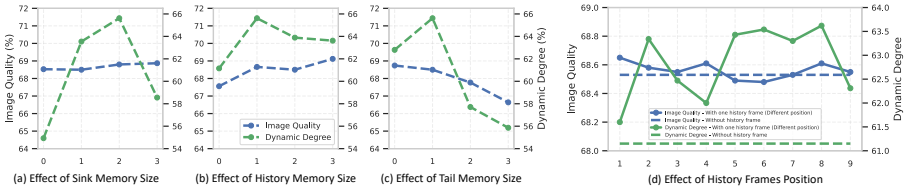

Fig. 2: Analysis of temporal memory in long-video extrapolation. (a-c): Varying the number of Sink, History, and Tail frames shows that increasing conditioning memory size does not consistently improve generation quality and may constrain motion dynamics. (d): Under a fixed memory budget, selecting history frames from different temporal positions leads to noticeable variations in motion dynamics while leaving image quality largely unchanged.

inference [4, 5, 7, 8, 12, 15, 35, 36, 38, 39, 41, 42, 47]. For example, Deep Forcing [36] compresses candidate memories based on query-averaged attention scores, while Reward Forcing [22] preserves full historical information through an exponential moving average mechanism. Other approaches explore memory compression to summarise long video histories into compact contexts [42, 47]. Despite these advances, existing studies largely treat memory as a capacity or update problem, leaving the temporal roles of different historical frames insufficiently understood. In particular, it remains unclear how much past information should be preserved and which parts of the history are most beneficial for long-horizon extrapolation.

Through empirical analysis of temporal memory configurations, we find that dense memory is not uniformly beneficial. Instead, its effectiveness depends critically on both the quantity and the temporal placement of stored frames. First, as shown in Fig. 2 (a-c), increasing the number of frames for conditioning, whether from Sink, middle-range History, or Tail, does not consistently improve long-horizon generation. Beyond a certain point, additional memory introduces redundancy that weakens motion dynamics without yielding meaningful gains in visual quality. Second, the temporal placement of historical frames also plays an important role. As shown in Fig. 2 (d), mid-range history contributes differently depending on where it is sampled, suggesting that not all past context is equally useful for future generation. Together, these findings indicate that temporal memory should not be treated merely as a chronological buffer whose effectiveness depends only on its size. Although existing approaches often employ memory compression or selection mechanisms, they primarily focus on improving scalability or retaining informative tokens, while the functional roles of different temporal regions remain underexplored. Simply retaining more past frames, or selecting historical context without considering their temporal roles, may even hinder long-horizon extrapolation. Instead, effective memory design must account for both how much past information is preserved and where it originates in the temporal sequence. This naturally calls for a structured view of temporal memory, where past frames play distinct functional roles rather than contributing uniformly.

Motivated by this insight, we propose Relaxed KV Memory, a structured sparse memory mechanism for AR diffusion, as shown in Fig. 3. Instead of employing the full or dense memory sequence as conditions, our approach decomposes temporal context into three functional roles: Sink for long-term stability,

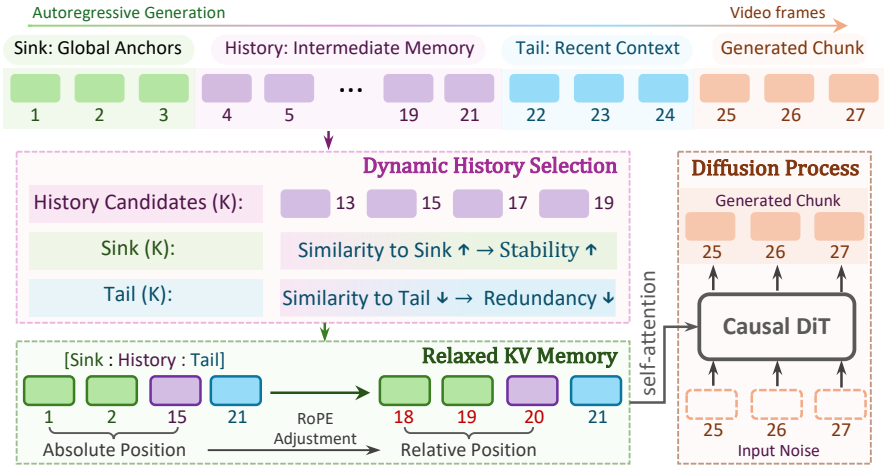

Fig. 3: Overview of Relaxed KV Memory. Instead of retaining dense chronological history, temporal memory is decomposed into three functional components: *Sink* for global anchors, *History* for intermediate motion structure, and *Tail* for recent continuity. During generation, candidate historical frames are dynamically selected to remain aligned with Sink while avoiding redundancy with Tail. The selected memory is then integrated through a relaxed KV formulation with adjusted relative positional encoding, enabling the model to leverage non-contiguous temporal context while preserving long-range consistency during autoregressive rollout.

Tail for short-term continuity, and dynamically selected History for mid-range structure, and selectively incorporates only the most relevant historical frames during generation. In particular, History frames are selected using a relaxation-based scoring mechanism that balances global stability against redundancy with recent context. This enables the model to prioritise informative mid-range frames while suppressing overly correlated historical signals. Consequently, the model avoids over-reliance on recent context while preserving globally consistent yet complementary temporal cues, mitigating error accumulation during extrapolation. By replacing rigid chronological conditioning with structured temporal roles, Relaxed KV Memory balances stability and motion evolution, improving long-horizon generation quality while reducing attention overhead.

In summary, the main contributions of this work are as follows:

- We conduct a systematic study of how temporal memory influences long-video extrapolation in AR diffusion. Our analysis identifies dense historical conditioning as a key bottleneck that limits long-horizon motion evolution.
- We propose Relaxed KV Memory, a sparse mechanism that dynamically selects relevant temporal context instead of preserving full history. This design reduces over-reliance on rigid past frames and enables more flexible long-term generation.
- Our approach significantly improves motion dynamics in long-video generation on VBench-Long, achieving a 66.8% improvement in Dynamic Degree and a 5.7% gain in the overall score, while reducing attention overhead and improving scalability.

2 Related Work

From Bidirectional to Autoregressive Video Diffusion. Recent advances in video generation have been largely driven by diffusion models that denoise all frames simultaneously using bidirectional attention, enabling strong visual fidelity but requiring access to the full sequence during generation [1, 17, 31, 33]. This design limits scalability to real-time or long-horizon scenarios where future frames are unavailable. To address this, a growing body of work has explored autoregressive (AR) formulations of video diffusion, enabling sequential generation that aligns with the causal structure of temporal data [18, 29, 37]. Early attempts reformulate diffusion training using teacher-forcing-style next-frame prediction [9, 45] or noise-independent training schemes such as diffusion forcing [3] and FAR [10]. Other methods also introduce queue-based conditioning and asynchronous generation strategies to extend temporal rollout, including FIFO-style history replacement [16], asynchronous AR diffusion [29], and masked autoregressive training [45]. History-aware conditioning [27], fast AR conversion from bidirectional models [37], and industrial-scale AR systems [18, 30] further demonstrate the feasibility of long-form sequential generation. However, these approaches inherently introduce a training–inference mismatch: models are typically trained with clean or distribution-smoothed historical context but must condition on their own imperfect predictions during autoregressive rollout. This mismatch leads to error accumulation over time, commonly referred to as exposure bias, motivating the development of rollout-aware training paradigms.

Mitigating Exposure Bias in AR Video Diffusion. To bridge the training-inference gap, recent work increasingly moves beyond ground-truth conditioning toward training regimes that explicitly model autoregressive dynamics. Self Forcing [13] unrolls generation during training and supervises the model on sequences conditioned on its own predictions, significantly reducing exposure bias. Then Self-Forcing++ [6] extends this idea to longer horizons, enabling minute-scale video synthesis with improved temporal stability. Other methods explore dynamic context updates through rolling conditioning strategies [21] or error recycling mechanisms for infinite-length generation [20]. Backward aggregation mitigates long-term degradation by incorporating historical corrections [23], while reward-guided methods introduce external supervision during rollout to steer generation toward higher-quality trajectories [22, 40]. End-to-end AR optimisation via self-resampling further aligns training with inference-time dynamics [11]. Despite these advances, exposure-bias mitigation alone does not fully resolve long-horizon degradation. As generation horizons extend, stability and efficiency increasingly depend on how historical information is stored, prioritised, and reused during self-attention.

Memory Management for Long Video Extrapolation. When extending AR diffusion to longer time horizons, drift becomes tightly coupled with memory usage. Existing approaches broadly fall into two categories: memory compression and memory optimisation. Memory compression methods aim to reduce attention overhead while preserving informative historical signals. Frame context packing reorganises past frames to stabilise next-frame prediction [41], while

KV cache compression strategies such as PackCache improve efficiency without retraining [19]. Pretraining objectives that preserve salient frames further enhance robustness under compressed memory [42]. Other approaches introduce structured tokens or adaptive memory flow to maintain long-term consistency under constrained attention budgets, including deep sink compression [36], adaptive memory flow mechanisms [15], and memorize-and-generate frameworks for real-time consistency [47]. In contrast, memory optimisation approaches treat historical context as a dynamic state to be refined during generation. Test-time adaptation strategies update model behaviour online to improve long-form synthesis [8], while hybrid state-space memory offers an alternative to standard KV attention for modelling long-range dependencies [39]. Salience-aware cache policies further prioritise informative tokens to improve temporal stability under limited memory budgets [4]. However, existing methods predominantly regulate how much memory is retained—through compression or adaptive updates—without explicitly considering how different temporal segments influence generation dynamics. In dense KV attention, recent context tends to dominate historical signals, which can amplify accumulated errors and lead to progressive temporal degradation during long-horizon extrapolation. In contrast, we argue that long-term stability depends not only on the quantity of stored memory, but on how temporal memory is utilised. By analysing the distinct functional roles of early anchors, mid-range history, and recent frames, we propose a relaxed KV memory design that selectively preserves informative temporal signals while mitigating error propagation. This structured utilisation of memory improves temporal consistency while reducing attention overhead.

3 Methods

3.1 Preliminaries

Autoregressive Video Diffusion. Given a video sequence $\{x^1, ..., x^N\}$, autoregressive (AR) video diffusion models the joint distribution as $p_\theta(x^{1:N}) = \prod_{i=1}^{N} p_\theta(x^i \mid c^{<i})$, where each conditional distribution is parameterised by a diffusion process that iteratively denoises a noisy latent z_t^i into the target frame x^i conditioned on historical context $c^{<i}$. The fundamental distinction among AR training paradigms lies in the formulation of this context. Under Teacher Forcing [9], the model conditions on pristine ground-truth history $c^{<i} = x^{<i} \sim P_{data}$, enabling parallelised training but inducing severe exposure bias, as inference requires conditioning on accumulated model predictions $\hat{x}^{<i} \sim P_\theta$ rather than clean observations. Diffusion Forcing [3] partially mitigates this mismatch by injecting continuous noise into the ground-truth context, defining $c^{<i} = z_t^{<i} \sim q(z_t^{<i}|x^{<i})$, yet isotropic Gaussian perturbations fail to replicate the structured sequential artifacts produced during actual autoregressive rollout. Self Forcing [13] achieves strict train-test alignment by directly defining $c^{<i} = \hat{x}^{<i} \sim P_\theta$, unrolling generation during training such that the denoising objective for each target frame x^i is optimised conditioned on frames sampled through the model's own reverse diffusion trajectory, forcing the network to directly observe and correct the structured artifacts it encounters at inference.

Long Video Rollout and Extrapolation. To synthesise sequences exceeding the fixed temporal capacity of the bidirectional teacher window, Self Forcing employs a chunk-wise sliding window strategy. Given a maximum window size L partitioned into chunks of length U, the model progressively accumulates generated chunks within a window, expanding $c^{<i} = \hat{x}^{<i} \sim P_\theta$ until the capacity limit L is reached. Upon saturation, the subsequent window is initialised by truncating the global context and conditioning solely on the final chunk of the preceding window as an overlapping temporal anchor $\hat{x}^{L-U:L} \sim P_\theta$, from which progressive chunk accumulation resumes. However, this rigid reliance on a fixed previously generated anchor forces residual errors and temporal biases to compound across successive window transitions, rendering long-horizon extrapolation susceptible to gradual semantic drift and overly constrained motion dynamics, highlighting the fundamental limitations of static memory conditioning.

Rotary Position Embedding. To manage positional information during sliding window extrapolation, Self Forcing employs a strictly localised RoPE [28] strategy that deliberately discards global temporal progression. Upon each window transition, the overlapping anchor $\hat{x}^{L-U:L}$ has its positional indices forcibly reset to $[0, U-1]$, with newly synthesised frames assigned local indices $[U, L-1]$, rendering the model entirely agnostic to the total number of frames generated across prior rollouts. This localised coordinate mapping coincides with periodic KV cache overwriting at each window boundary, compelling the model to treat every extrapolation step as an independent short-horizon task. However, this rigid consecutive index reassignment fundamentally destroys true relative temporal distances between frames, artificially compressing temporal gaps and forcing historically distant frames to appear adjacent to recent ones. Consequently, this inflexible RoPE mapping structurally prevents the model from leveraging long-range historical context, severely limiting its capacity to preserve global visual coherence and accurate motion dynamics over extended durations.

3.2 Temporal Memory Analysis for Long Video Extrapolation

Despite the robustness gained from self-forcing training, long-horizon extrapolation still relies on repeatedly conditioning generation on previously synthesised frames. Under the sliding-window inference paradigm described above, these historical frames are preserved as KV memory and reused across successive rollout steps. Consequently, the effectiveness of long video generation critically depends on how this temporal memory is constructed and utilised. Existing autoregressive diffusion methods typically adopt a dense memory design, retaining all available past frames within the attention window. This implicitly assumes that historical context contributes uniformly to future generations. However, our empirical findings in Sec. 1 suggest that this assumption does not hold.

Sensitivity to Memory Quantity and Temporal Placement. As shown in Fig. 2 (a-c), increasing the number of Sink, History, or Tail frames for conditioning does not consistently improve generation quality. While visual fidelity remains relatively stable, motion dynamics exhibit a non-monotonic trend as memory size grows. In particular, excessive memory often constrains motion evolution rather than enhancing temporal coherence, indicating that dense conditioning may introduce redundancy instead of useful guidance. Furthermore,

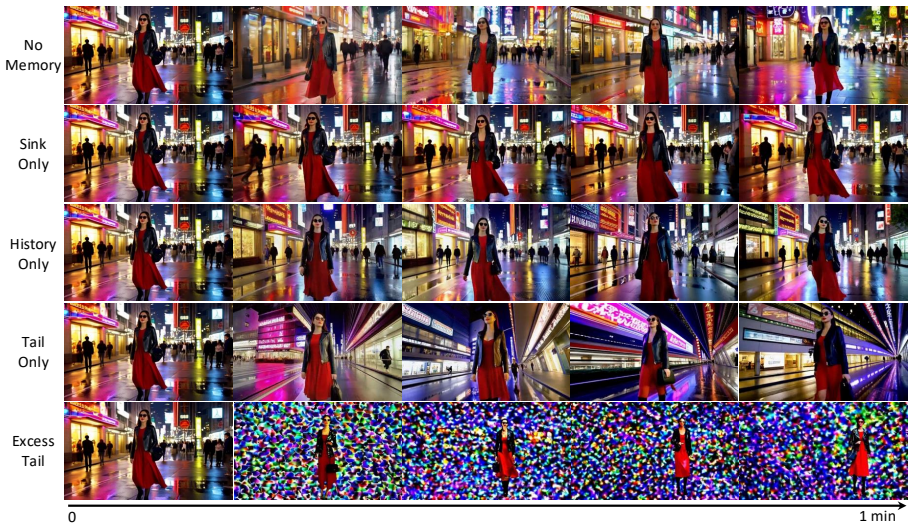

Fig. 4: Temporal memory roles during long-horizon rollout. Different memory configurations lead to distinct failure modes over time. Without memory, generation becomes fragmented. Sink-only conditioning preserves appearance but limits motion evolution. History-only conditioning enables motion variation but weakens identity consistency. Tail-only conditioning improves short-term continuity but destabilises long-term rollout, while excessive Tail leads to generation collapse. These results highlight the heterogeneous roles of temporal memory in balancing stability and motion dynamics.

Fig. 2 (d) demonstrates that the temporal placement of historical frames plays a critical role. Under a fixed memory budget, selecting history frames from different temporal locations leads to noticeable variations in motion dynamics, while visual consistency and image quality remain largely unchanged. This suggests that historical context is not interchangeable and primarily influences structural motion rather than appearance stability.

Functional Roles and Interdependence of Temporal Regions. To further interpret these observations, we conduct qualitative analyses by isolating different memory components during long-horizon rollout, as shown in Fig. 4. Distinct failure modes emerge depending on which component is used. Without memory, generation becomes fragmented over time. Sink-only conditioning preserves appearance but limits motion evolution, resulting in repetition. Tail-only conditioning improves short-term continuity but destabilises long-term rollout, while excessive Tail leads to collapse. History-only conditioning enables motion variation but weakens identity consistency.

We further analyse the interaction between these components by selectively removing each from a balanced configuration, as shown in Fig. 5. Removing the sink causes temporal drift. Removing History limits motion dynamics. Removing the tail weakens short-term continuity and results in rigid motion. Only a balanced combination of Sink, History, and Tail maintains both stability and dynamic evolution.

Taken together, these findings indicate that temporal memory is inherently heterogeneous and should not be treated as a homogeneous chronological buffer.

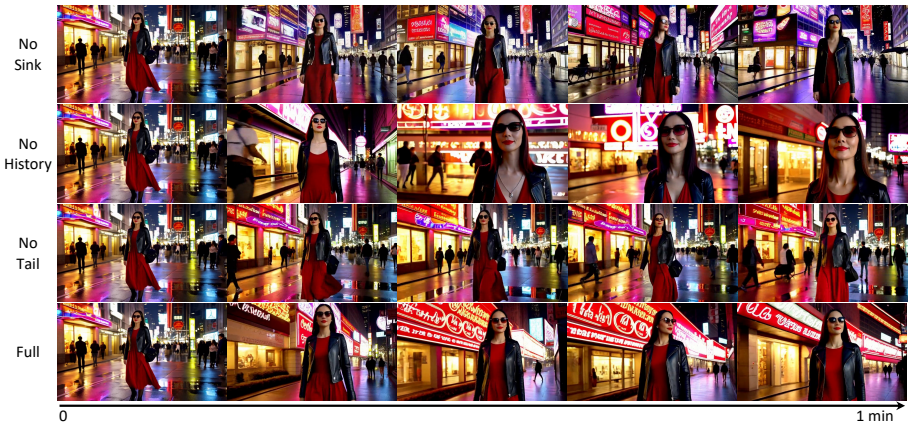

Fig. 5: Complementary roles of temporal memory components during long-horizon rollout. Removing individual components leads to distinct failure modes. Without Sink, the generation exhibits temporal drift. Without History, motion evolution becomes limited and repetitive. Without Tail, short-term continuity is weakened, resulting in rigid motion. A balanced combination of Sink, History, and Tail maintains both stability and dynamic evolution.

Instead, effective long-horizon generation requires balancing complementary functional roles: *early frames provide global anchors, recent frames maintain short-term continuity, and mid-range frames contribute structural motion cues*. This motivates a structured formulation of temporal memory, which we introduce next.

3.3 Relaxed KV Memory

KV Memory Selection. Relaxed KV Memory decomposes temporal conditioning into three functional components: a fixed Sink $\mathcal{S}$ for long-term anchors, a step-dependent Tail $\mathcal{T}_i$ ensuring short-term continuity at generation step i, and a dynamically selected History set $\mathcal{H}_i$ drawn from an intermediate candidate region to stabilise and contextualise the current prediction.

At generation step i, the available historical frames are partitioned as

$$\hat{x}^{<i} = \mathcal{S} \cup \mathcal{M}_i^{cand} \cup \mathcal{T}_i, \tag{1}$$

where $\mathcal{M}_i^{cand}$ denotes the mid-range candidate frames. Based on the empirical observation that informative history frames tend to appear in the later part of the middle region, as illustrated in Fig. 2 (d), we restrict candidate selection to the second half of $\mathcal{M}_i^{cand}$. Formally, we define a reduced candidate set

$$\tilde{\mathcal{M}}_i = \left\{ h \in \mathcal{M}_i^{cand} \mid \mathrm{idx}(h) \geq \tfrac{1}{2} |\mathcal{M}_i^{cand}| \right\}, \tag{2}$$

where $\mathrm{idx}(h)$ denotes the temporal index of frame h within $\mathcal{M}_i^{cand}$. For each candidate frame $h \in \tilde{\mathcal{M}}_i$, we compute a representative key prototype

$$\tilde{K}_h = \mathrm{norm}\left(\frac{1}{|\Omega(h)|} \sum_{k \in \Omega(h)} K_k \right). \tag{3}$$

where $\Omega(h)$ denotes the set of tokens belonging to frame h. Similarly, aggregated Sink and Tail prototypes are defined as $\tilde{K}_{\mathcal{S}} = \mathrm{norm}\big(\mathcal{K}(\mathcal{S})\big)$ and $\tilde{K}_{\mathcal{T}_i} = \mathrm{norm}\big(\mathcal{K}(\mathcal{T}_i)\big)$.

We then compute a stability score and a redundancy score $S(h) = \tilde{K}_h^\top \tilde{K}_{\mathcal{S}}$ and $R(h) = \tilde{K}_h^\top \tilde{K}_{\mathcal{T}_i}$. The relaxation score of candidate frame h is defined as: $r(h) = S(h) - \lambda R(h)$, where λ controls the trade-off between global stability and local redundancy.

Finally, History frames are selected via Top-K ranking

$$\mathcal{H}_i = \mathrm{TopK}\big(\{r(h)\}_{h \in \tilde{\mathcal{M}}_i}\big), \tag{4}$$

and the memory used for conditioning is constructed as $\mathcal{M}_i = \mathcal{S} \cup \mathcal{H}_i \cup \mathcal{T}_i$.

Extrapolation With Sparse Memory. In standard sliding-window inference, conditioning relies on a dense contiguous buffer of the most recent frames: $c^{<i} = \hat{x}^{i-L:i}$, where L denotes the attention window size.

In contrast, Relaxed KV replaces this dense chronological buffer with a structured temporal memory

$$c^{<i} = \mathcal{M}_i = \mathcal{S} \cup \mathcal{H}_i \cup \mathcal{T}_i, \tag{5}$$

where the Sink $\mathcal{S}$ provides a fixed global anchor, while the History $\mathcal{H}_i$ and Tail $\mathcal{T}_i$ are dynamically selected during generation. Specifically, at each rollout step the model constructs $\mathcal{M}_i$ on the fly by selecting informative mid-range history frames according to the relaxation score $r(h)$, while always retaining the Sink anchors and the most recent Tail frames. This design is motivated by our empirical observations in Fig. 4, which show that excess memory often introduces redundant context and constrains motion evolution. By retaining only a small set of informative history frames together with global anchors and recent context, Relaxed KV avoids repeatedly conditioning on redundant past frames. The corresponding attention keys are therefore constructed as

$$\mathbf{K}^i_{attn} = \mathcal{K}(\mathcal{M}_i). \tag{6}$$

As a result, sparse memory conditioning mitigates error amplification during long-horizon rollout while preserving both global consistency and evolving motion dynamics.

Position Embedding Adjustment. Relaxed KV Memory constructs a non-contiguous conditioning set $\mathcal{M}_i = \mathcal{S} \cup \mathcal{H}_i \cup \mathcal{T}_i$, which makes standard sliding-window RoPE resetting unsuitable. We therefore adopt a hybrid positional indexing scheme that treats recent Tail and distant (Sink or History) memory differently.

For Tail, we preserve its real temporal location by applying RoPE with absolute frame indices. Let i denote the current generation frame index and let $|\mathcal{T}_i|$ be the number of Tail frames. We define the Tail start index as $p_{\mathcal{T}} = i - |\mathcal{T}_i|$, and apply RoPE to Tail keys using indices $\{p_{\mathcal{T}}, \ldots, i-1\}$. For Sink and History, instead of using their original absolute positions, we apply RoPE with relative indices anchored immediately before the Tail segment. Let $|\mathcal{S}|$ and $|\mathcal{H}_i|$ denote

the number of Sink and History frames, and define $p_{\mathcal{SH}} = p_{\mathcal{T}} - (|\mathcal{S}| + |\mathcal{H}_i|)$. We then assign Sink plus History indices to the contiguous range $\{p_{\mathcal{SH}}, \ldots, p_{\mathcal{T}} - 1\}$ and apply RoPE accordingly.

This adjustment enforces a consistent ordering of memory roles in positional space, Sink or History preceding Tail, without forcing all memory to be locally contiguous within the sliding window. By keeping Tail on absolute positions while anchoring Sink or History relatively to the current context, the model can leverage long-range anchors and mid-range structure without collapsing them into recent dynamics during window shifts.

4 Experiments

4.1 Experimental Settings

Implementation Details. We implement chunk-wise Self Forcing [13] as our base model. Videos are generated in a chunk-wise autoregressive manner with a chunk size of 3 frames. For Relaxed Memory, we adopt the following hyper-parameter settings: the number of sink frames is set to 2, the number of tail frames is set to 1, and one history frame is selected from a candidate pool of size 4. The coefficient λ is set to 2.0. All methods are evaluated under the same inference settings for fair comparison. For throughput evaluation (FPS), following previous work, all experiments are conducted on a single NVIDIA H100 GPU.

Evaluation. We evaluate long-horizon video generation using the VBench-Long benchmark [14], following the protocol adopted in recent autoregressive video diffusion works [6, 36]. Specifically, we use 128 prompts from MovieGen [24], which are commonly used for long video evaluation. Following prior work [13, 22,36], each prompt is refined using Qwen2.5-7B-Instruct [25] to improve prompt clarity and diversity. We generate long videos with durations of 30s and 60s and evaluate them using the VBench-Long metrics, including subject consistency, background consistency, aesthetic quality, imaging quality, motion smoothness, and dynamic degree. In addition, we report CLIP-based temporal metrics to quantify drift and repetition across frames. Further implementation details are provided in the *Appendix*.

4.2 Comparisons to State of the Art

We compare Relax Forcing with recent autoregressive video diffusion methods on VBench-Long under both 30-second and 60-second settings. As shown in Tab. 1, Relax Forcing achieves the highest overall score in both regimes. For 30-second videos, our method reaches 80.87%, outperforming the strongest training-free baseline, Deep Forcing, by +0.93%. Under the more challenging 60-second setting, Relax Forcing achieves the result of 80.88%, while other methods show noticeable degradation, indicating improved robustness for long-horizon generation. The most significant gain is observed in Dynamic Degree, reflecting stronger motion evolution over long sequences. This suggests that structured temporal memory selection effectively mitigates the over-constrained generation caused by dense chronological memory accumulation. Importantly, the improved motion dynamics do not compromise visual fidelity: Subject and Background Consistency remain competitive, and Imaging Quality is comparable to prior

Table 1: Quantitative comparison on long video generation on VBench-Long.

Methods	Training Free	Throughput (FPS)	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Dynamic Degree	Average
			30 seconds						
CausVid [37] [CVPR'25]	×	15.78	97.92	96.77	59.77	66.36	98.08	47.21	77.69
Self Forcing [13] [NeurIPS'25]	×	15.78	97.34	96.47	59.44	68.58	98.63	36.62	76.18
Rolling Forcing [21] [ICLR'26]	×	15.75	**98.07**	**96.84**	60.75	**70.73**	98.74	32.71	76.31
LongLive [32] [ICLR'26]	×	**18.16**	97.97	96.83	**61.51**	69.07	**98.76**	45.55	78.28
Deep Forcing [36] [arXiv]	√	15.79	97.34	96.48	60.68	69.31	98.27	57.56	79.94
Relax Forcing (Ours)	√	16.33	96.99	96.12	60.13	68.50	97.80	**65.67**	**80.87**
			60 seconds						
CausVid [37] [CVPR'25]	×	15.78	97.81	96.75	59.42	65.84	98.09	46.44	77.39
Self Forcing [13] [NeurIPS'25]	×	15.78	96.31	**96.82**	56.45	66.33	98.21	31.98	74.35
Rolling Forcing [21] [ICLR'26]	×	15.75	**97.94**	96.76	60.02	**70.72**	98.71	32.50	76.11
LongLive [32] [ICLR'26]	×	**18.16**	97.85	96.74	**61.29**	69.11	**98.75**	43.49	77.87
Deep Forcing [36] [arXiv]	√	15.79	96.96	96.32	59.86	69.27	98.23	57.19	79.64
Relax Forcing (Ours)	√	16.33	96.81	95.97	59.58	68.66	97.74	**66.49**	**80.88**

Table 2: Effect of different memory conditioning strategies.

Methods	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Dynamic Degree	Average
Baseline (Self Forcing)	97.22	96.39	59.18	68.43	**98.39**	39.38	76.50
+ Full Attention	**97.25**	**96.60**	59.22	66.43	98.29	45.03	77.14
+ Attention Sink	97.18	96.40	59.80	**68.58**	98.32	54.37	79.11
+ **Relaxed KV Memory (Ours)**	96.99	96.12	**60.13**	68.50	97.80	**65.67**	**80.87**

methods. Moreover, since Relax Forcing operates on a sparse structured memory rather than the full dense history, it reduces the amount of self-attention computation, leading to a modest improvement in inference throughput compared with previous methods.

Among the compared methods, Self Forcing [13] reduces exposure bias through self-rollout, Rolling Forcing [21] improves robustness via noise scheduling during rollout, and LongLive [32] enhances long-context modelling through architectural refinements. These approaches rely on modified training procedures to improve rollout stability. In contrast, Relax Forcing introduces no additional training objective and instead improves long-horizon generation by organising inference-time memory with structured sparsity.

4.3 Ablation Analysis

We conduct ablation studies on the 30-second setting to validate key design choices in Relax Forcing, including the structured memory design, the history selection strategy, the candidate pool size, and the robustness to the redundancy weight λ. Additional qualitative comparisons and video demonstrations for ablations are provided in the *supplementary material*.

Effect of Structured Memory Design. We first analyse how different memory conditioning strategies influence long-horizon generation. Starting from the baseline Self Forcing model, we progressively introduce three designs: full attention over all historical frames, sink anchoring, and the proposed relaxed memory selection. As shown in Tab. 2, enabling full attention slightly improves motion dynamics compared with the baseline, indicating that additional historical context can provide useful temporal cues. Introducing sink anchoring further improves Dynamic Degree, highlighting the importance of maintaining stable global anchors during long-horizon rollout. Finally, our relaxed memory design significantly boosts Dynamic Degree while preserving visual consistency. This

Table 3: Comparison of different history selection strategies in Relaxed KV Memory.

Methods	Imaging Quality ↑	Dynamic Degree ↑	**Overall ↑**	Drift ↓	Repetition ↓	**Balance ↓**
Baseline (Self Forcing)	68.58	36.62	76.18	2.70	**81.34**	1.000
Random Sampling	68.55	62.47	80.42	2.15	83.59	0.805
Fixed-Positional Sampling	68.51	64.22	80.59	2.18	83.16	0.768
Attention-Based Sampling	68.80	63.18	80.54	2.17	83.20	0.765
First-Half Sampling	**68.83**	59.76	80.27	**1.68**	87.87	1.000
Relax Forcing (Ours)	68.50	**65.67**	**80.87**	2.13	83.33	**0.745**

Table 4: Effect of the size of candidate history frames used for Relaxed KV selection.

Candidates Number	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Dynamic Degree	Average
2	96.98	612	60.16	68.52	97.79	65.55	80.85
3	**97.00**	96.12	**60.17**	**68.57**	**97.80**	65.39	80.84
4	96.99	96.13	60.12	68.50	**97.80**	**65.62**	**80.86**
5	**97.00**	**96.14**	60.00	68.48	97.79	65.30	80.78
7	96.94	96.09	60.14	68.46	97.77	63.44	80.47
9	96.96	96.09	60.04	68.38	97.76	63.97	80.53

demonstrates that selectively retaining informative history frames is more effective than relying on dense chronological memory.

Effectiveness of History Selection Strategy. We evaluate several strategies for selecting History frames from the candidate memory pool, including random sampling, fixed-positional sampling, attention-based sampling, and selecting frames from the first half of the candidate region. As shown in Tab. 3, all variants achieve similar imaging quality, but differ noticeably in motion dynamics. Relax Forcing achieves the highest Dynamic Degree at 65.67%, and the best overall score at 80.87%, outperforming all alternative history selection strategies. Furthermore, we observe an inherent trade-off between temporal drift and repetition. For instance, First-Half Sampling minimises Drift but suffers from severe Repetition, whereas the Baseline exhibits the exact opposite behaviour. To appropriately quantify this trade-off, we introduce a normalised Balance score, defined as the sum of min-max scaled Drift and Repetition. Our Relax Forcing achieves the lowest Balance score of 0.745, indicating an optimal equilibrium that ensures stable generation without excessive repetition or temporal drift.

Sensitivity to Candidate Pool Size. We further study how the size of the candidate history pool influences performance. As shown in Tab. 4, the overall performance remains stable when the candidate pool contains a small number of frames (2–4). However, increasing the pool size beyond this range gradually reduces Dynamic Degree and the overall Average score.

This behaviour can be explained by the way candidate history frames are constructed. To improve efficiency, candidate frames are uniformly sampled from the second half of the middle region. As the number of candidates increases, the sampled frames become increasingly redundant in temporal content, introducing overlapping or conflicting motion cues. This redundancy weakens the benefit of sparse memory selection and increases the risk of propagating accumulated errors during long-horizon rollout. In contrast, a moderate candidate pool provides sufficient diversity for selecting informative history frames while avoiding

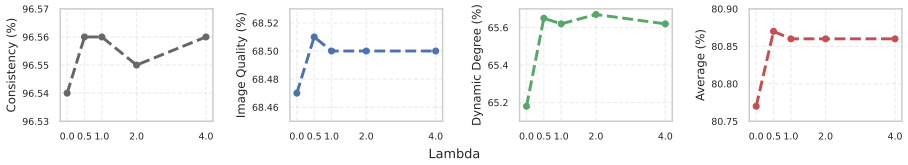

Fig. 6: Sensitivity to the redundancy weight λ in Relaxed KV Memory. Varying λ leads to minor changes in visual consistency and image quality, while moderately influencing motion dynamics. Overall performance remains stable across a wide range of values, indicating that Relaxed KV is robust to the stability–redundancy trade-off.

excessive redundancy. In our experiments, using four candidate frames achieves the best balance between efficiency and motion evolution.

Robustness to Redundancy Weight λ. Finally, we analyse the sensitivity of Relaxed KV Memory to the redundancy weight λ, which balances stability from Sink alignment and redundancy suppression with Tail frames. As illustrated in Fig. 6, varying λ across a wide range leads to only minor changes in visual consistency and image quality, while moderately influencing motion dynamics. The overall Average score remains stable across different values of λ, demonstrating that Relax Forcing is not sensitive to precise hyperparameter tuning. This robustness indicates that the proposed relaxation scoring mechanism provides a stable trade-off between preserving global anchors and encouraging diverse motion evolution.

4.4 Human Evaluation

To further assess the perceptual quality of the generated videos, we conduct a user preference study comparing Relax Forcing with prior methods. Participants are asked to evaluate pairs of videos based on visual quality (VQ), motion quality (MQ), and text–video alignment (TA). Additional implementation details are provided in the Appendix.

Table 5: Human preference (%).

Methods	VQ ↑	MQ ↑	TA ↑	AVG ↑
Self Forcing	3.1	6.2	2.3	3.9
Attention Sink	9.2	6.9	13.8	10.0
Rolling Forcing	43.1	21.5	32.3	32.3
Relax Forcing	**44.6**	**65.4**	**51.5**	**53.8**

As shown in Tab. 5, Relax Forcing is consistently preferred over prior methods across all evaluation criteria. In particular, our method receives significantly higher preference in motion quality and average quality, indicating that the proposed structured memory mechanism improves long-horizon motion dynamics while maintaining competitive visual fidelity.

5 Conclusion

In this work, we investigate the role of temporal memory in autoregressive video diffusion for long-horizon generation. We show that the main limitation of minute-scale synthesis lies not in memory capacity but in how historical context is utilised during inference. Based on this insight, we propose Relaxed KV Memory, which decomposes temporal context into Sink, Tail, and dynamically selected History to enable structured and sparse conditioning. Experiments on VBench-Long demonstrate improved motion dynamics and temporal consistency while reducing attention overhead.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 2, 5
2. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: ICML (2024) 2
3. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. In: NeurIPS. vol. 37, pp. 24081–24125 (2024) 2, 5, 6
4. Chen, H., Xu, C., Yang, X., Chen, X., Deng, C.: Past-and future-informed kv cache policy with salience estimation in autoregressive video diffusion. arXiv preprint arXiv:2601.21896 (2026) 3, 6
5. Chen, S., Wei, C., Sun, S., Nie, P., Zhou, K., Zhang, G., Yang, M.H., Chen, W.: Context forcing: Consistent autoregressive video generation with long context. arXiv preprint arXiv:2602.06028 (2026) 2, 3
6. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025) 2, 5, 11, 20
7. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Lol: Longer than longer, scaling video generation to hour. arXiv preprint arXiv:2601.16914 (2026) 3
8. Dalal, K., Koceja, D., Xu, J., Zhao, Y., Han, S., Cheung, K.C., Kautz, J., Choi, Y., Sun, Y., Wang, X.: One-minute video generation with test-time training. In: CVPR. pp. 17702–17711 (2025) 3, 6
9. Gao, K., Shi, J., Zhang, H., Wang, C., Xiao, J., Chen, L.: Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. arXiv preprint arXiv:2411.16375 (2024) 5, 6
10. Gu, Y., Mao, W., Shou, M.Z.: Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325 (2025) 5
11. Guo, Y., Yang, C., He, H., Zhao, Y., Wei, M., Yang, Z., Huang, W., Lin, D.: End-to-end training for autoregressive video diffusion via self-resampling. arXiv preprint arXiv:2512.15702 (2025) 5
12. Hong, Y., Mei, Y., Ge, C., Xu, Y., Zhou, Y., Bi, S., Hold-Geoffroy, Y., Roberts, M., Fisher, M., Shechtman, E., et al.: Relic: Interactive video world model with long-horizon memory. arXiv preprint arXiv:2512.04040 (2025) 2, 3
13. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: NeurIPS (2025) 2, 5, 6, 11, 12, 19
14. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., Wang, Y., Chen, X., Chen, Y.C., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench++: Comprehensive and versatile benchmark suite for video generative models. IEEE TPAMI (2025) 11, 20
15. Ji, S., Chen, X., Yang, S., Tao, X., Wan, P., Zhao, H.: Memflow: Flowing adaptive memory for consistent and efficient long video narratives. arXiv preprint arXiv:2512.14699 (2025) 3, 6
16. Kim, J., Kang, J., Choi, J., Han, B.: Fifo-diffusion: Generating infinite videos from text without training. NeurIPS pp. 89834–89868 (2024) 5

17. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 2, 5
18. Li, D., Fei, Z., Li, T., Dou, Y., Chen, Z., Yang, J., Fan, M., Xu, J., Wang, J., Gu, B., et al.: Skyreels-v3 technique report. arXiv preprint arXiv:2601.17323 (2026) 5
19. Li, K., Shah, M., Shang, Y.: Packcache: A training-free acceleration method for unified autoregressive video generation via compact kv-cache. arXiv preprint arXiv:2601.04359 (2026) 6
20. Li, W., Pan, W., Luan, P.C., Gao, Y., Alahi, A.: Stable video infinity: Infinite-length video generation with error recycling. In: ICLR (2025) 5
21. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025) 2, 5, 12
22. Lu, Y., Zeng, Y., Li, H., Ouyang, H., Wang, Q., Cheng, K.L., Zhu, J., Cao, H., Zhang, Z., Zhu, X., et al.: Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. arXiv preprint arXiv:2512.04678 (2025) 2, 3, 5, 11
23. Po, R., Chan, E.R., Chen, C., Wetzstein, G.: Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. arXiv preprint arXiv:2512.12080 (2025) 2, 5
24. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024) 11, 20
25. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025) 11, 20
26. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Shechtman, E., Huang, X.: Motionstream: Real-time video generation with interactive motion controls. arXiv preprint arXiv:2511.01266 (2025) 2
27. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. In: ICML (2025) 5
28. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 7, 19
29. Sun, M., Wang, W., Li, G., Liu, J., Sun, J., Feng, W., Lao, S., Zhou, S., He, Q., Liu, J.: Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In: CVPR. pp. 7364–7373 (2025) 2, 5
30. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025) 5
31. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 **3**(4), 6 (2025) 2, 5
32. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. In: ICLR (2025) 2, 12
33. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024) 2, 5

34. Ye, D., Zhou, F., Lv, J., Ma, J., Zhang, J., Lv, J., Li, J., Deng, M., Yang, M., Fu, Q., et al.: Yan: Foundational interactive video generation. arXiv preprint arXiv:2508.08601 (2025) 2
35. Yesiltepe, H., Meral, T.H.S., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. arXiv preprint arXiv:2511.20649 (2025) 3
36. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025) 3, 6, 11, 12, 20
37. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: CVPR. pp. 22963–22974 (2025) 2, 5, 12
38. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: SIGGRAPH Asia. pp. 1–11 (2025) 2, 3
39. Yu, Y., Wu, X., Hu, X., Hu, T., Sun, Y., Lyu, X., Wang, B., Ma, L., Ma, Y., Wang, Z., et al.: Videossm: Autoregressive long video generation with hybrid state-space memory. arXiv preprint arXiv:2512.04519 (2025) 3, 6
40. Zhang, J., Li, N., Ban, Y., Bai, A., Cui, J.: Reward-forcing: Autoregressive video generation with reward feedback. arXiv preprint arXiv:2601.16933 (2026) 2, 5
41. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. In: NeurIPS (2025) 3, 5
42. Zhang, L., Cai, S., Li, M., Zeng, C., Lu, B., Rao, A., Han, S., Wetzstein, G., Agrawala, M.: Pretraining frame preservation in autoregressive video memory compression. arXiv preprint arXiv:2512.23851 (2025) 3, 6
43. Zhao, M., He, G., Chen, Y., Zhu, H., Li, C., Zhu, J.: Riflex: A free lunch for length extrapolation in video diffusion transformers. In: ICML (2025) 2
44. Zhao, M., Zhu, H., Wang, Y., Yan, B., Zhang, J., He, G., Yang, L., Li, C., Zhu, J.: Ultravico: Breaking extrapolation limits in video diffusion transformers. arXiv preprint arXiv:2511.20123 (2025) 2
45. Zhou, D., Sun, Q., Peng, Y., Yan, K., Dong, R., Wang, D., Ge, Z., Duan, N., Zhang, X.: Taming teacher forcing for masked autoregressive video generation. In: CVPR. pp. 7374–7384 (2025) 5
46. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026) 2
47. Zhu, T., Zhang, S., Sun, Z., Tian, J., Tang, Y.: Memorize-and-generate: Towards long-term consistency in real-time video generation. arXiv preprint arXiv:2512.18741 (2025) 3, 6

Appendix

This appendix provides supplementary material that supports the main paper. The contents are organised as follows:

- **A: Algorithm Details.** A detailed description of the proposed Relaxed KV Memory selection procedure.
- **B: Experimental Settings.** Implementation details and inference configurations used in our experiments.
- **C: Memory Size Analysis.** Extended results when varying the number of Sink, History, and Tail frames.
- **D: Latency Analysis.** Runtime breakout of our method.
- **E: Human Evaluation Details.** Additional information about the user study protocol and evaluation setup.
- **F: Qualitative Comparisons.** Additional visual comparisons with representative methods, including CausVid, Self-Forcing, Attention Sink, Rolling Forcing, and Deep Forcing.

A Algorithm Details

A.1 Relaxed KV Memory Selection

Algorithm 1 summarises the proposed Relaxed KV Memory selection procedure. At each rollout step, the available historical frames are decomposed into three parts: a fixed Sink $\mathcal{S}$, a step-dependent Tail $\mathcal{T}_i$, and an intermediate candidate region $\mathcal{M}_i^{cand}$. Based on the observation in Fig. 2(d) of the main paper, we restrict candidate history selection to the latter half of $\mathcal{M}_i^{cand}$ and rank candidate frames using the proposed relaxation score, which balances alignment with Sink against redundancy with Tail.

Algorithm 1: Relaxed KV Memory Selection

Input: Generated frames $\hat{x}^{<i}$, Sink size N_S, Tail size N_T, History size N_H, redundancy weight λ

Output: structured memory $\mathcal{M}_i$

$\mathcal{S} \leftarrow \hat{x}^{1:N_S}$, $\mathcal{T}_i \leftarrow \hat{x}^{i-N_T:i}$, $\mathcal{M}_i^{\text{cand}} \leftarrow \hat{x}^{N_S+1:i-N_T-1}$

$\tilde{\mathcal{M}}_i \leftarrow \{h \in \mathcal{M}_i^{\text{cand}} \mid \text{idx}(h) \geq \frac{1}{2}|\mathcal{M}_i^{\text{cand}}|\}$

$\tilde{K}_{\mathcal{S}} \leftarrow \text{norm}(\mathcal{K}(\mathcal{S}))$, $\tilde{K}_{\mathcal{T}_i} \leftarrow \text{norm}(\mathcal{K}(\mathcal{T}_i))$

foreach $h \in \tilde{\mathcal{M}}_i$ **do**

$\tilde{K}_h \leftarrow \text{norm}\Big(\frac{1}{|\Omega(h)|}\sum_{k\in\Omega(h)} K_k\Big)$

$r(h) \leftarrow \tilde{K}_h^\top \tilde{K}_{\mathcal{S}} - \lambda\, \tilde{K}_h^\top \tilde{K}_{\mathcal{T}_i}$

$\mathcal{H}_i \leftarrow \text{TopK}_{N_H}\big(\{r(h)\}_{h\in\tilde{\mathcal{M}}_i}\big)$

$\mathcal{M}_i \leftarrow \mathcal{S} \cup \mathcal{H}_i \cup \mathcal{T}_i$

return $\mathcal{M}_i$

A.2 Sparse Extrapolation During Rollout

In the standard sliding-window inference of Self Forcing [13], conditioning relies on a dense contiguous buffer of the most recent frames,

$$c^{<i} = \hat{x}^{i-L:i},$$

where L is the attention window size. In contrast, Relaxed KV replaces this dense memory with a structured memory

$$c^{<i} = \mathcal{M}_i = \mathcal{S} \cup \mathcal{H}_i \cup \mathcal{T}_i.$$

At each rollout step, $\mathcal{H}_i$ is selected on the fly according to the relaxation score, while $\mathcal{S}$ is kept fixed and $\mathcal{T}_i$ is updated using the most recent frames. This design avoids repeatedly conditioning on redundant recent history and mitigates error amplification during long-horizon extrapolation.

A.3 Hybrid Positional Indexing

Because $\mathcal{M}_i$ is non-contiguous in time, standard sliding-window RoPE [28] resetting is unsuitable. We therefore apply RoPE differently to Tail and Sink/History memory:

- **Tail:** absolute frame indices are preserved so that the most recent context remains temporally grounded.
- **Sink and History:** relative indices are assigned immediately before the Tail segment, preserving their role ordering without forcing them to be mapped as temporally adjacent to the latest frames.

This hybrid indexing allows the model to exploit both global anchors and mid-range structure without collapsing them into short-term dynamics.

B Additional Experimental Settings

B.1 Inference Configuration

Unless otherwise specified, all experiments use chunk-wise Self Forcing [13] as the base model with chunk size $U = 3$ frames. For Relaxed KV Memory, we use:

- Sink frames: 2
- Tail frames: 1
- History frames: 1
- Candidate pool size: 4
- Redundancy weight λ: 2.0

All methods are evaluated under the same inference setting for fair comparison.

B.2 Prompt Set and Evaluation Protocol

Following prior work on long-video autoregressive diffusion [6,36], we evaluate on 128 prompts from MovieGen [24]. To improve prompt clarity and diversity, the prompts are refined using Qwen2.5-7B-Instruct [25]. For robustness evaluation, we generate five videos for each prompt.

We report the standard VBench-Long [14] metrics for both 30-second and 60-second generation. The evaluated metrics include:

- **Subject Consistency.** Measures whether the main subject maintains consistent appearance throughout the video. This metric is computed by measuring DINO feature similarity across frames.
- **Background Consistency.** Evaluates the temporal consistency of background scenes by computing CLIP feature similarity across frames.
- **Aesthetic Quality.** Assesses the visual aesthetics of generated frames using the LAION aesthetic predictor, which correlates with human perception of composition, color harmony, realism, and overall artistic quality.
- **Imaging Quality.** Measures image-level distortions such as blur, noise, or exposure artifacts. This metric is computed using the MUSIQ image quality predictor trained on the SPAQ dataset.
- **Motion Smoothness.** Evaluates the temporal smoothness of motion in generated videos using motion priors derived from a video frame interpolation model.
- **Dynamic Degree.** Quantifies the magnitude of motion in generated videos. Optical flow estimated by RAFT is used to measure the degree of dynamic movement across frames.

For ablation experiments, we additionally report two complementary temporal metrics:

- **Drift.** Measures long-range semantic drift by computing the CLIP feature difference between the first and last 5-second clips of the generated video.
- **Repetition.** Measures temporal repetition using the average CLIP2CLIP similarity among clips within the same generated video.
- **Balance.** A normalised combination of Drift and Repetition that reflects the trade-off between temporal drift and excessive repetition. Lower values indicate a better balance.

C Extended Analysis on Memory Quantity

We provide additional analysis by independently varying the number of Sink, History, and Tail frames while keeping the remaining components fixed, as shown in Tab. C.1. These experiments help understand how different temporal roles influence long-horizon generation.

Effect of Sink Frames. Sink frames serve as global anchors that provide long-term identity and scene stability. Increasing the number of Sink frames initially improves generation quality. When the number of Sink frames increases from 0 to 2, the overall score improves significantly (78.36% $\rightarrow$ 80.86%), mainly due

Table C.1: Extended analysis of memory allocation. We independently vary the number of Sink, History, and Tail frames while keeping the other components fixed.

Component	Size	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Dynamic Degree	Overall
Sink	0	95.84	95.26	58.92	67.56	97.68	54.93	78.36
	1	96.85	95.98	59.29	68.66	97.80	63.54	80.36
	2	96.99	96.13	60.12	68.50	97.80	65.62	**80.86**
	3	97.46	96.46	60.87	69.12	98.13	58.55	80.10
History	0	96.98	96.10	60.33	68.53	97.81	61.15	80.15
	1	96.99	96.13	60.12	68.50	97.80	65.62	**80.86**
	2	97.12	96.21	60.20	68.80	97.95	63.89	80.69
	3	97.19	96.27	60.17	68.87	98.00	63.62	80.69
Tail	0	96.62	96.01	60.24	68.74	96.96	64.39	80.49
	1	96.99	96.13	60.12	68.50	97.80	65.62	**80.86**
	2	97.21	96.31	60.38	67.77	98.04	57.71	79.57
	3	97.17	96.36	60.26	66.64	98.03	55.86	79.05

to improved Dynamic Degree and stronger visual consistency. However, adding more Sink frames begins to over-constrain the generation process. With three Sink frames, Dynamic Degree drops sharply (65.62% → 58.55%), indicating that excessive global anchors may restrict motion evolution.

Effect of Tail Frames. Tail frames provide short-term temporal continuity during autoregressive rollout. Using a single Tail frame achieves the best balance between stability and motion dynamics. Increasing the number of Tail frames slightly improves short-term consistency metrics such as Subject and Background Consistency, but significantly reduces Dynamic Degree (65.62% → 55.86% when Tail increases from 1 to 3). This suggests that excessive reliance on recent frames may overly constrain motion evolution.

Effect of History Frames. History frames capture mid-range temporal context that supports motion progression. Introducing a single History frame substantially improves Dynamic Degree compared with the no-history setting (61.15% → 65.62%), leading to the best overall performance. However, increasing the number of History frames beyond one does not provide additional benefits. Larger History sets introduce redundant temporal information and slightly degrade Dynamic Degree and the overall score.

Effect of History Frame Position. We further analyse how the temporal position of the selected History frame influences generation performance while fixing the memory configuration to Sink=2 and Tail=1, as shown in Tab. C.2.

Overall, visual quality metrics such as Subject Consistency, Background Consistency, Aesthetic Quality, and Imaging Quality remain relatively stable across different history positions. In contrast, motion-related metrics show noticeable variation. Compared with the no-history setting, introducing a history frame generally improves Dynamic Degree, confirming the importance of mid-range temporal context for long-horizon extrapolation.

We further observe that positions located in the middle portion of the candidate region tend to achieve slightly higher Dynamic Degree and better overall scores, whereas earlier or later positions provide weaker improvements. This ob-

Table C.2: Extended analysis on history frame selection positions while fixing the memory configuration to Sink=2 and Tail=1.

History Position	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Dynamic Degree	Overall
None	96.98	96.10	60.33	68.53	97.81	61.15	80.15
0	97.12	96.22	60.66	68.65	97.88	61.60	80.36
1	96.99	96.13	60.41	68.58	97.80	63.34	80.54
2	97.06	96.18	60.44	68.55	97.85	62.47	80.42
3	97.07	96.19	60.46	68.61	97.85	62.00	80.36
4	96.93	96.08	60.15	68.49	97.76	63.43	80.47
5	96.99	96.12	60.14	68.48	97.79	63.54	80.51
6	96.96	96.12	60.02	68.53	97.80	63.30	80.45
7	96.93	96.09	60.02	68.61	97.76	63.62	80.51
8	97.03	96.18	60.04	68.55	97.83	62.31	80.32

Table D.1: Latency profiling comparison between the dense baseline and Relax Forcing. The baseline performs self-attention over 21 frames, while Relax Forcing reduces the attention length to 7 frames using structured KV memory.

Metric	Baseline (21f)	Relax Forcing (7f)	Speedup
Flash Attention	444.2 ms	168.1 ms	2.64×
KV Select + RoPE	125.5 ms	114.1 ms	1.10×
Candidate Scoring	0.0 ms	3.4 ms	–
KV Cache Management	13.0 ms	49.5 ms	0.26×
Total Self-Attention	678.5 ms	430.8 ms	1.58×
Block Generation Time	848 ms	627 ms	1.35×
Diffusion Time	67654 ms	49515 ms	1.37×
End-to-End Time	89122 ms	70650 ms	1.26×

servation supports our design choice of selecting informative mid-range history frames rather than relying on fixed chronological memory.

Summary. Overall, these results suggest that temporal memory should be carefully balanced rather than simply expanded. The best configuration in our experiments is obtained with Sink=2, History=1, and Tail=1, which achieves the highest overall score while maintaining strong motion dynamics.

D Latency Analysis

We further analyse the inference efficiency of Relax Forcing by profiling its runtime against a dense-memory baseline under identical generation settings. All experiments generate a one-minute video on a single NVIDIA H100 GPU.

Baseline configuration. The baseline uses dense self-attention over 18 historical frames together with the current 3-frame block, resulting in an effective self-attention length of 21 frames.

Relax Forcing configuration. Our method adopts the structured KV memory design with 2 Sink frames, 1 History frame, and 1 Tail frame. Together with the current 3-frame block, this results in an effective self-attention length of 7 frames. Although Relax Forcing introduces an additional candidate-scoring step to select

the History frame, it substantially reduces the number of tokens participating in attention.

The results in Tab. D.1 show that Relax Forcing significantly reduces the cost of the dominant attention operations. In particular, flash attention is reduced from 444.2 ms to 168.1 ms per block, corresponding to a 2.64× speedup, which closely matches the reduction in effective attention length (21 frames → 7 frames).

Although Relax Forcing introduces a candidate-scoring step for History frame selection, the overhead is extremely small (3.4 ms per block). In the full generation process, this accounts for only about 0.8% of the diffusion runtime, making it negligible compared with the attention savings.

The profiling also reveals a moderate increase in KV cache management time, which rises from 13.0 ms to 49.5 ms due to the larger candidate region maintained in the cache. However, this additional cost is outweighed by the substantial reduction in flash-attention computation.

Overall, Relax Forcing reduces the accumulated self-attention time by 1.58×, leading to a 1.37× speedup in diffusion generation and a 1.26× end-to-end acceleration. These results confirm that structured KV memory improves not only long-horizon generation quality but also inference efficiency.

E Human Evaluation Details

We conduct a user preference study to further assess perceptual quality. Participants are shown generated videos from Relax Forcing and comparison methods and are asked to select the preferred result under the following criteria:

- **Visual Quality (VQ):** sharpness, realism, and overall visual plausibility
- **Motion Quality (MQ):** temporal smoothness, natural motion evolution, and absence of drift or repetition
- **Text–Video Alignment (TA):** consistency between the generated video and the text prompt

For each prompt, videos from different methods are presented in randomised order to avoid position bias. Each comparison is independently annotated by multiple participants, and the final preference score is computed as the percentage of pairwise wins. We compare Relax Forcing against Self Forcing, Attention Sink, and Rolling Forcing, as reported in Tab. 5 of the main paper. We additionally provide the webpage interface as shown in Fig. E.1.

F Additional Qualitative Comparisons

We provide additional qualitative comparisons with representative methods, including Self Forcing, Attention Sink, Rolling Forcing, and our Relax Forcing.

These qualitative results highlight three recurring patterns:

Temporal Drift. Methods with insufficient long-range anchoring often exhibit gradual semantic drift, including changes in identity, scene layout, or object appearance over time.

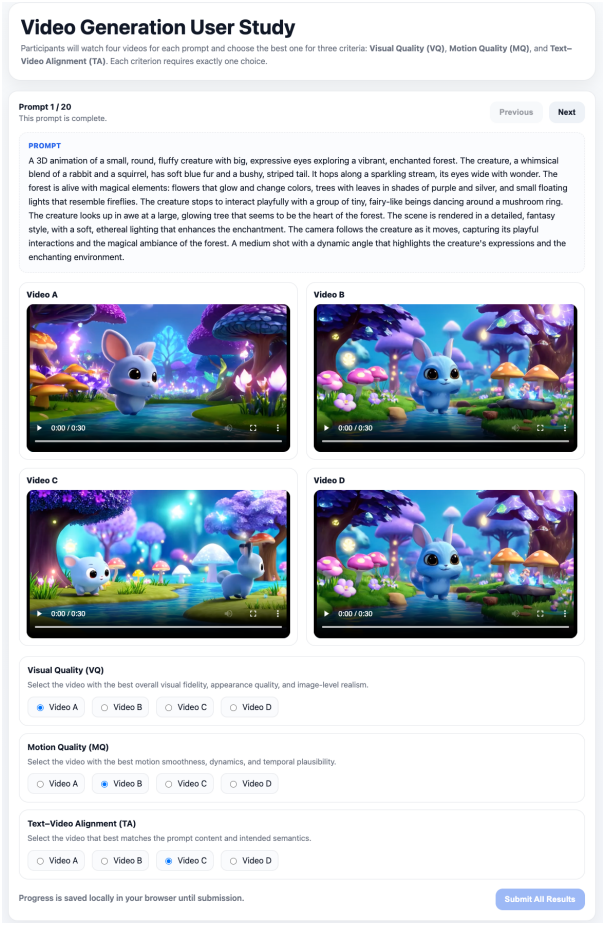

Fig. E.1: User study interface. For each prompt, participants watch four generated videos from different methods and select the best one according to three criteria: visual quality (VQ), motion quality (MQ), and text–video alignment (TA).

Over-Constrained Motion. Methods that rely heavily on recent context or dense chronological memory may maintain a stable appearance but produce repetitive or rigid motion.

Balanced Evolution. Relax Forcing preserves global consistency while allowing sustained motion evolution, producing videos that remain coherent without collapsing into either drift or repetition.

We also include additional video demos in the supplementary material to illustrate these temporal differences more clearly than static frames alone.

Prompt: A charming comic-style illustration depicting a cozy living room scene where a fluffy gray cat is waking up its sleeping owner, who lies on the couch with a sleepy, resigned expression. The cat, with large, round eyes and a mischievous look, is pawing at the owner's face and meowing insistently. The owner attempts to ignore the cat, turning away slightly, but the cat persists, jumping onto the owner's chest and nuzzling their hand. Finally, the owner, unable to resist, reaches under the pillow and pulls out a small bag of treats, offering it to the cat with a playful smile. The background shows soft, warm lighting from a nearby lamp, with scattered books and a blanket on the couch. A medium shot from a slightly elevated angle, capturing both the cat and the owner's interaction.

Fig. F.1: Qualitative comparison of long video generation under different methods.

Prompt: A stunning mid-afternoon landscape photograph with a low camera angle, showcasing several giant wooly mammoths treading through a snowy meadow. Their long, wooly fur gently billows in the brisk wind as they move, creating a sense of natural movement. Snow-covered trees and dramatic snow-capped mountains loom in the distance, adding to the majestic setting. Wispy clouds and a high sun cast a warm glow over the scene, enhancing the serene and awe-inspiring atmosphere. The depth of field brings out the detailed textures of the mammoths and the snowy environment, capturing every nuance of these prehistoric giants in breathtaking clarity.

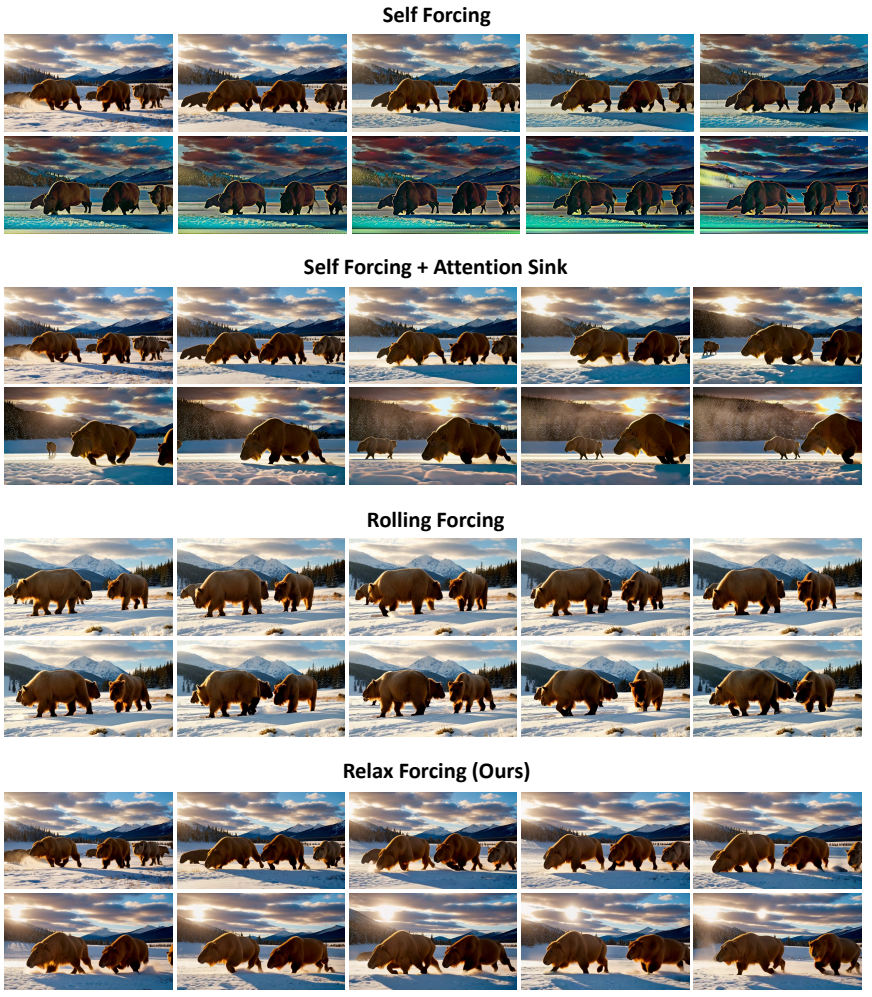

Fig. F.2: Qualitative comparison of long video generation under different methods.

Prompt: A stylish woman strolls down a bustling Tokyo street, the warm glow of neon lights and animated city signs casting vibrant reflections. She wears a sleek black leather jacket paired with a flowing red dress and black boots, her black purse slung over her shoulder. Sunglasses perched on her nose and a bold red lipstick add to her confident, casual demeanor. The street is damp and reflective, creating a mirror-like effect that enhances the colorful lights and shadows. Pedestrians move about, adding to the lively atmosphere. The scene is captured in a dynamic medium shot with the woman walking slightly to one side, highlighting her graceful strides.

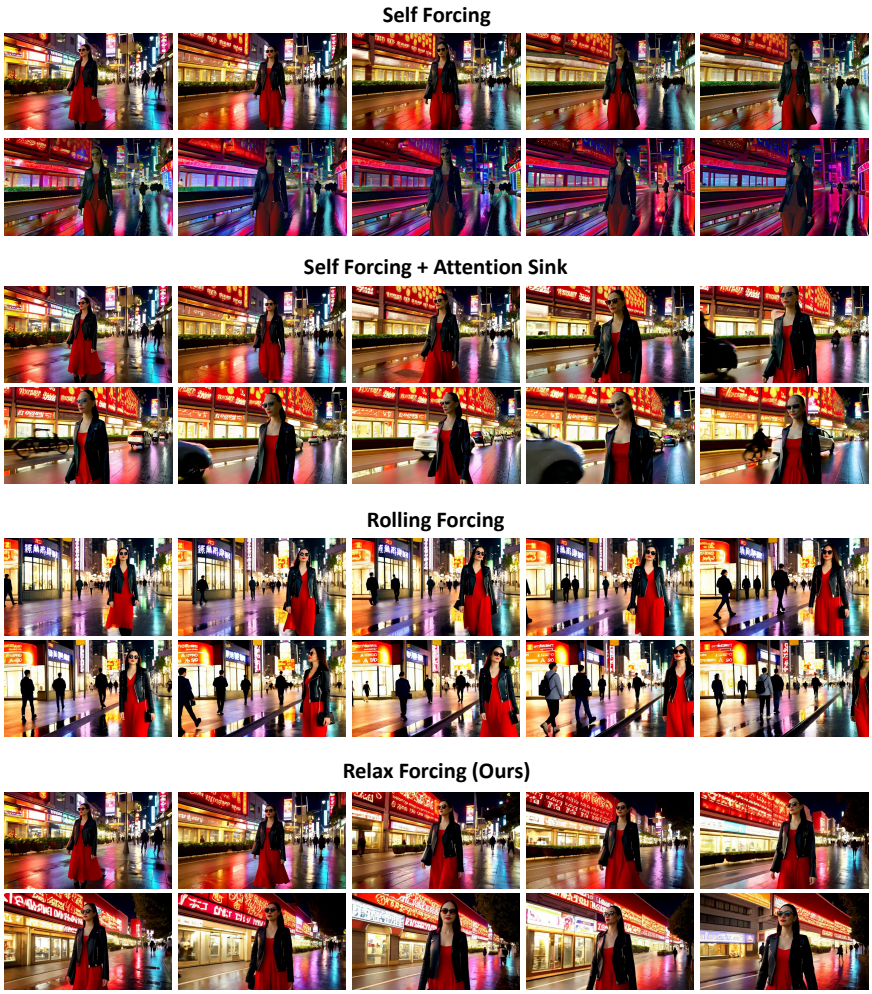

Fig. F.3: Qualitative comparison of long video generation under different methods.

Prompt: A drone view of waves crashing against the rugged cliffs along Big Sur's Garay Point beach. The crashing blue waters create white-tipped waves, while the golden light of the setting sun illuminates the rocky shore, casting long shadows. In the distance, a small island with a lighthouse stands tall, its beam piercing the twilight. Green shrubbery covers the cliff's edge, and the steep drop from the road down to the beach is a dramatic feat, with the cliff's edges jutting out over the sea. The camera angle provides a bird's-eye view, capturing the raw beauty of the coast and the rugged landscape of the Pacific Coast Highway. The scene is bathed in a warm, golden hue, highlighting the textures and details of the rocky terrain.

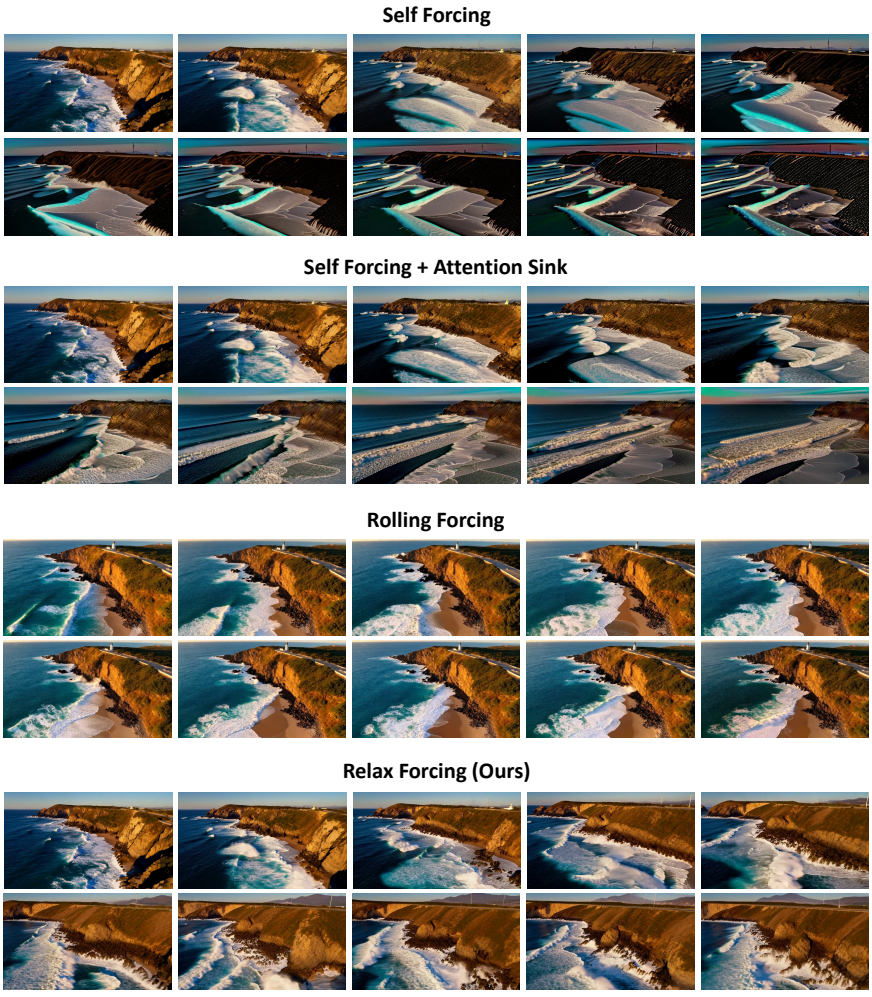

Fig. F.4: Qualitative comparison of long video generation under different methods.

Prompt: A dynamic and vibrant anime illustration in a flowing watercolor style, capturing the bustling snowy streets of Tokyo. The camera moves smoothly through the city, following several people joyfully enjoying the snow and shopping at nearby stalls. Gorgeous sakura petals dance through the air, swirling with snowflakes. The scene features traditional Japanese architecture, with shops and lanterns illuminated by the soft winter light. People are bundled up in warm coats and scarves, their faces lit with smiles. The background shows blurred, snowy rooftops and distant cherry blossom trees, creating a serene yet lively atmosphere. A medium shot with a sweeping camera motion, highlighting the natural movement of both people and petals.

Fig. F.5: Qualitative comparison of long video generation under different methods.