MEMACTIVE: TRAINING-FREE MEMORY ACTIVATION FOR CONSISTENT INTERACTIVE STREAMING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

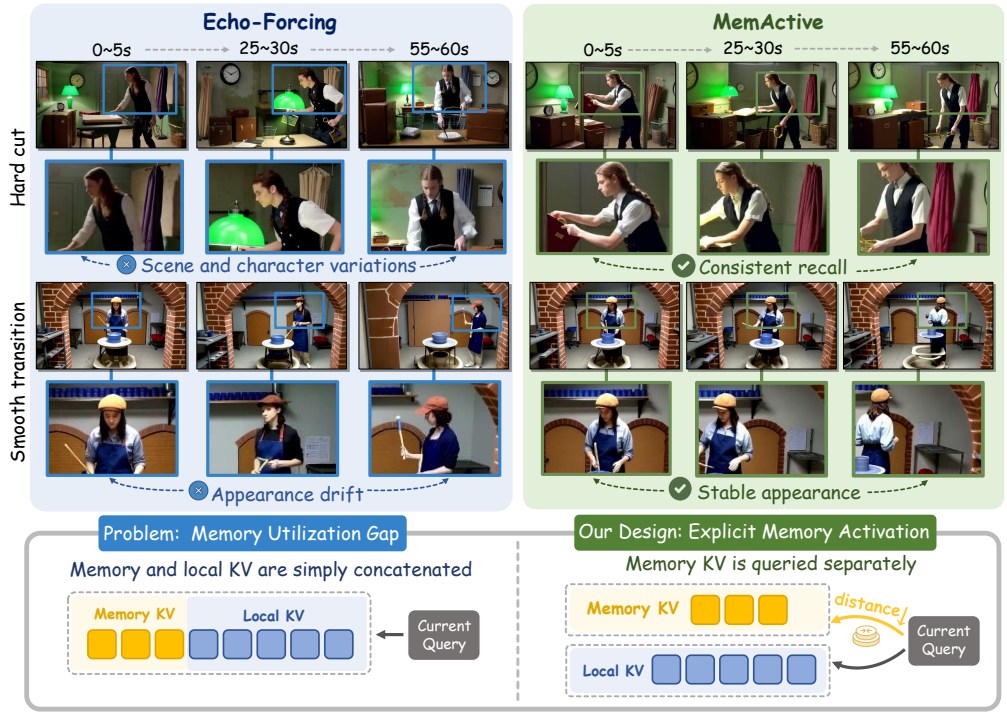


Figure 1: **MemActive enables consistent long-term recall.** Directly concatenated historical KV states can remain underused, leading to inconsistent entity recall after hard cuts and appearance drift during smooth transitions. MemActive adjusts query positions along the memory route while retaining native local positions.

ABSTRACT

Interactive streaming video generation requires following evolving prompts while recalling entities such as characters or scenes beyond the recent context. However, frozen generators can fail to recover visual content even when relevant historical memory is supplied. We term this failure the *memory-utilization gap*. In a controlled experiment with fixed memory, temporal RoPE compression improves consistency with the supplied history, accompanied by an increase in memory attention mass. Motivated by this finding, we introduce **MemActive**, a training-free framework combining Route-Specific Query RoPE (RSQ-RoPE) with Semantic Structured Memory (SSM). RSQ-RoPE adjusts memory-route positions according to entity-description compatibility while retaining native local-route positions. SSM supplies entity-specific memory for retrieval and recombination across prompts. Quantitative and qualitative experiments under hard cuts and smooth transitions show that MemActive preserves recurring content over long temporal gaps. These consistency gains further extend to four additional training and training-free streaming backbones.

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.

1 INTRODUCTION

Recent autoregressive video diffusion models (Huang et al., 2025; Zhu et al., 2026; Liu et al., 2026b) combine causal attention with KV caching to enable real-time interactive streaming (Yang et al., 2026a;b; Wu et al., 2026), where a video evolves with successive instructions. Interactive prompts require recurring entities to retain their identity beyond the recent context while following requested changes in action, appearance, or scene.

Existing approaches address long-term consistency by constructing and retrieving historical memory (Yang et al., 2026a; Hu et al., 2026; Ji et al., 2025; Wu et al., 2026; Kim et al., 2026; Li et al., 2026). However, when retrieved memory is directly prepended to the recent context, its influence on subsequent generation can remain limited, especially in training-free settings where the frozen generator has not been trained to condition on such memory. As shown in Figure 1, memory-augmented methods such as Echo-Forcing (Wu et al., 2026) still lose recurring entities after hard cuts and exhibit appearance drift during smooth transitions.

To examine memory use independently of retrieval, we supply a fixed memory anchor from an earlier clip. Under native attention, this memory receives little attention and its visual content is poorly recovered. Compressing the query–memory RoPE offset (Su et al., 2024) improves recall, identifying positional compatibility as a factor in the *memory-utilization gap*. Historical memory provides recurring entity information, while recent context supports local temporal continuity. The challenge is to improve memory utilization without disturbing native query–local-context relations.

To address this challenge, we introduce **MemActive**, a training-free framework combining **Route-Specific Query RoPE (RSQ-RoPE)** with **Semantic Structured Memory (SSM)**. RSQ-RoPE gives each query separate positional views for the two sources: the memory view adjusts its temporal offset according to entity-description compatibility, while the local view retains native positions. Both routes share one softmax distribution. SSM stores character and scene states separately, providing entity-addressable memories for direct retrieval and flexible recombination across prompts. Together, SSM determines *what memory is retrieved*, while RSQ-RoPE determines *how the retrieved memory is utilized* by adjusting query–memory positional compatibility.

Qualitative and quantitative experiments show that MemActiveimproves recurring-content consistency across hard cuts and smooth transitions while maintaining generation quality, with consistent gains across four additional streaming backbones.

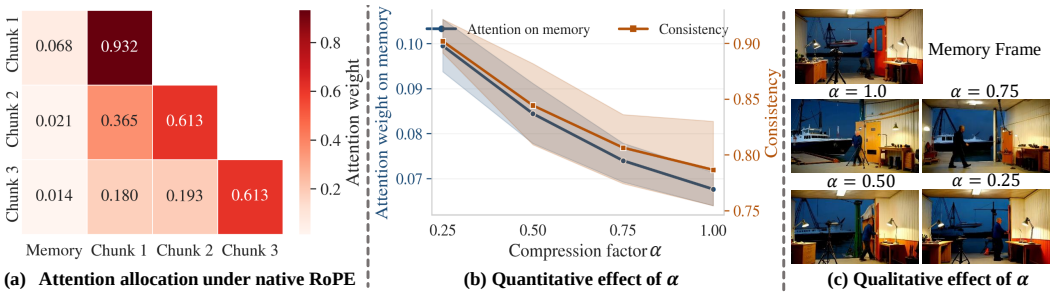
Our contributions are summarized as follows:

- We identify a *memory-utilization gap* in frozen streaming generation and show that temporal RoPE remapping improves recall in a controlled experiment with supplied memory.
- We propose **MemActive**, a training-free framework in which SSM constructs and retrieves entity-addressable memory, while RSQ-RoPE adjusts query–memory positional relations without altering the native local-route mapping.
- Experiments demonstrate improved consistency under hard cuts and smooth transitions, with gains extending to four additional streaming backbones.

2 RELATED WORK

Autoregressive Streaming Video Generation. Long-video diffusion extends generation through noise rescheduling (Qiu et al., 2024) or autoregressive prediction (Chen et al., 2024; Henschel et al., 2025). Recent work has improved training–inference alignment, generation efficiency, and long-horizon stability, making real-time streaming increasingly practical (Huang et al., 2025; Zhu et al., 2026; Liu et al., 2026b). These advances provide efficient backbones, but our prompt-interactive setting requires recalling history beyond the local window.

Memory for Prompt-Interactive Streaming Video Generation. Local-window KV caching supports efficient streaming by retaining only recent states (Huang et al., 2025; Yesiltepe et al., 2026), weakening long-term consistency across prompts. Recent methods (Yang et al., 2026a; Tan et al., 2026; Yang et al., 2026b; Wu et al., 2026; Kim et al., 2026; Bian et al., 2026; Zhao et al., 2026; Hu et al., 2026; Liu et al., 2026a; Xu et al., 2026) construct, retain, or retrieve historical memory to preserve long-term consistency. While MemFlow (Ji et al., 2025) and TetherMem (Li et al., 2026)



(a) Attention allocation under native RoPE **(b) Quantitative effect of α** **(c) Qualitative effect of α**

Figure 2: **Temporal RoPE compression improves recall with fixed supplied memory.** We provide a memory anchor from an earlier clip and generate a new clip under a different prompt. (a) Under native RoPE, attention is overwhelmingly concentrated on the current and preceding local chunks, with only a small fraction assigned to the supplied memory. (b) Across the tested α values, stronger compression improves anchor consistency, accompanied by increased memory attention. (c) As $\alpha \downarrow$, the generated character and background more closely match those in the memory anchor.

study query-aware selection and routing of historical memory in video generation, MemActive focuses on how a frozen generator effectively uses retrieved memory across successive prompts in prompt-interactive streaming.

RoPE in Long-Horizon Video Generation. RoPE (Su et al., 2024) encodes temporal relations, but generators (Wan et al., 2025) pretrained within a bounded temporal range can struggle with long rollouts. Prior methods (Yesiltepe et al., 2026; Yi et al., 2026; Kim et al., 2026; Tian et al., 2026; Yang et al., 2026b; Zhao et al., 2026) reindex cached positions or organize sink, history, and local-window caches for long-horizon generation. RSQ-RoPE instead constructs separate memory and local query routes, improving long-term visual consistency while maintaining motion smoothness.

3 Motivation

To examine whether a generator effectively uses historical memory, we supply a memory anchor from the previous clip while generating several chunks of the next clip, which depicts the same character and scene under a new prompt. We assess visual recall through quantitative and qualitative consistency with the memory anchor, and record attention allocation to characterize memory access. Across 100 cases, we hold the supplied memory and random seed fixed while varying only the query–memory temporal mapping. See Appendix D for details.

For a spatiotemporal token at position (t, w, h), where t denotes its temporal position and (w, h) its spatial position, let $R(t, w, h)$ denote the corresponding RoPE rotation (Su et al., 2024). Given pre-RoPE query and key vectors q_i and k_j, their self-attention logit is

$$\ell_{ij} = \frac{\left[R(t_i, w_i, h_i)q_i\right]^\top \left[R(t_j, w_j, h_j)k_j\right]}{\sqrt{d}}, \qquad a_{ij} = \mathrm{Softmax}_j(\ell_{ij}). \tag{1}$$

Here, i indexes the query token, j indexes the key token, and d denotes the attention head dimension. We focus on the temporal dimension and therefore omit the spatial coordinates (w, h), denoting the corresponding rotation by $R_T(t)$. Let t_c be the temporal position of the current query and t_m the temporal position assigned to the memory anchor. When the query reads the memory, we use the memory-route logit

$$\ell_{c,m}(\alpha) = \frac{\left[R_T(\tilde{t}_c)q_c\right]^\top \left[R_T(t_m)k_m\right]}{\sqrt{d}}, \qquad \tilde{t}_c = t_m + \alpha(t_c - t_m), \quad 0 < \alpha \leq 1. \tag{2}$$

Here, $\tilde{t}_c$ denotes the query position remapped toward the memory anchor. Thus, $\alpha = 1$ recovers the native temporal relation, whereas a smaller α compresses the query–memory temporal offset. We provide a detailed analysis of temporal-distance compression under RoPE in Appendix C.

Observation 1: Supplied memory is poorly recovered under native RoPE. Under native RoPE ($\alpha = 1$), consistency with the supplied anchor is low (Figure 2(b)), and the qualitative results

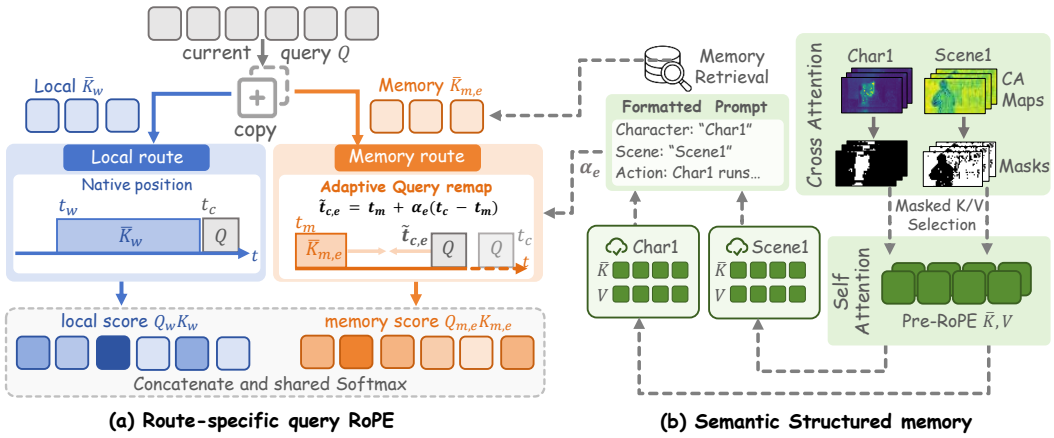


Figure 3: **Overview of MemActive.** (a) RSQ-RoPE keeps native temporal positions for the local route, applies entity-specific query remapping to compress the query–memory offset in the memory route, and combines both routes through a shared softmax. (b) SSM derives entity masks from structured prompts to store character- and scene-specific pre-RoPE KV states, then retrieves the requested memories for RSQ-RoPE.

show weaker preservation of its character and background (Figure 2(c)). In this setting, attention is concentrated on the current and preceding local chunks, with little mass assigned to memory (Figure 2(a)).

Observation 2: Temporal RoPE remapping improves recall. Across the tested α values, stronger compression improves consistency with the fixed anchor, accompanied by increased memory attention (Figure 2(b)). This intervention identifies the temporal mapping as a controllable factor in recall. Remapping changes individual memory logits as well as total attention mass (Appendix C), so the measured gain reflects the positional intervention as a whole.

4 Method

MemActive combines RSQ-RoPE for memory access with SSM for entity-memory construction and retrieval (Figure 3). For the current chunk, we denote the pre-RoPE query by Q, the cached pre-RoPE keys and corresponding values in the recent local window by $(\bar{K}_w, V_w)$, and those retrieved by SSM by $(\bar{K}_m, V_m)$.

4.1 Route-Specific Query RoPE

Section 3 shows that temporal RoPE compression improves consistency with the supplied memory. However, applying this remapping to local attention alters native query–local offsets and may disrupt temporal continuity. We therefore give each pre-RoPE query two positional views: native for the local route and compressed toward retrieved memory for the memory route.

Following existing cache layouts (Yesiltepe et al., 2026; Yi et al., 2026; Kim et al., 2026; Tian et al., 2026; Yang et al., 2026b), we arrange retrieved memory, the recent local window, and the current chunk sequentially on a shared temporal axis. All retrieved entity groups share the reindexed temporal anchor t_m, regardless of their original clip. Thus, $t_c - t_m$ denotes a cache-coordinate offset rather than elapsed video time. Let q be a pre-RoPE query at t_c, $\bar{k}_w$ a local key at t_w, and $\bar{k}_{m,e}$ a memory key for entity e. Spatial coordinates remain unchanged. The local route uses the native positions:

$$q_w = R_T(t_c)q, \qquad k_w = R_T(t_w)\bar{k}_w. \tag{3}$$

The memory route instead remaps the query position toward t_m:

$$\tilde{t}_{c,e} = t_m + \alpha_e(t_c - t_m), \qquad q_{m,e} = R_T(\tilde{t}_{c,e})q, \qquad k_{m,e} = R_T(t_m)\bar{k}_{m,e}, \quad 0 < \alpha_e \leq 1. \tag{4}$$

With $\alpha_e = 1$, the memory route recovers the native offset for the same memory and cache layout. Smaller values compress this offset. The same pre-RoPE query q receives a local view at t_c and

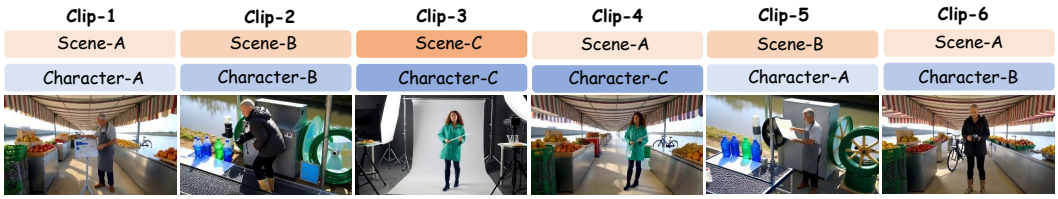


Figure 4: **Flexible character–scene composition with SSM.** Separate character and scene memories enable recombination across prompts.

an entity-specific memory view at $\tilde{t}_{c,e}$. Reusing the memory view for local keys would alter their native query–local offsets by a magnitude of $(1-\alpha_e)(t_c-t_m)$. Query-side remapping allows a fixed memory-key sequence to be reused across query positions. Appendix C shows that an equivalent key-side design requires query-dependent key rotations.

Stored entity appearances may differ from those requested by later prompts. We use description similarity as a proxy for compatibility. Let u_c and u_m denote the current and stored descriptions of entity e. We compute their CLIP text-feature cosine similarity and clip it to $[0, 1]$ to obtain s_e. We set

$$\alpha_e = \lambda + (1 - s_e)(1 - \lambda), \qquad 0 < \lambda \leq 1. \tag{5}$$

Higher compatibility gives stronger compression: $s_e = 1$ yields $\alpha_e = \lambda$, while $s_e = 0$ restores the native offset ($\alpha_e = 1$) without removing memory. Entity identifiers retrieve relevant memories, while description similarity controls query–memory offset compression to reduce stale-appearance interference. Let Q_w, K_w, $Q_{m,e}$, and $K_{m,e}$ denote the full query and key sequences for the local window and entity-specific memory groups, respectively. Their route-specific query–key scores are

$$L_w = \frac{Q_w K_w^\top}{\sqrt{d}}, \qquad L_{m,e} = \frac{Q_{m,e} K_{m,e}^\top}{\sqrt{d}}, \qquad L_m = \mathrm{Concat}_e(L_{m,e}). \tag{6}$$

We concatenate the local and memory scores along the key dimension and normalize them with a shared softmax:

$$[A_w \| A_m] = \mathrm{Softmax}([L_w \| L_m]), \tag{7}$$

where A_w and A_m are the normalized weights and $\|$ denotes concatenation along the key dimension. For fixed features, local logits remain unchanged, but shared normalization lets memory logits alter local attention mass. The corresponding attention output is $O = A_w V_w + A_m V_m$. For fixed query and key features, an entity-wise additive bias changes only the total attention assigned to each memory group while preserving the relative weights of tokens within the group. In contrast, RSQ-RoPE changes query–key positional compatibility and can therefore reweight individual memory tokens. We compare these mechanisms in Table 4.

4.2 Semantic Structured Memory

JSON

```
"clip 1": {
  "scene": <scene>,
  "characters": [{
    "name": <entity>,
    "description": <desc>
  }],
  "action": <action>
}
```

Figure 5: **Structured prompt.**

To support direct retrieval and flexible recombination across prompts, SSM organizes historical states into character- and scene-specific memories indexed by structured prompt fields.

Structured Prompt. Each clip is represented by a structured prompt containing its scene, active characters, and action, as illustrated in Figure 5. Structured semantic fields provide consistent addresses for entity memory across prompt changes. This structure enables flexible character–scene composition, as shown in Figure 4.

Entity-level Memory Construction and Retrieval. For each entity, SSM constructs a fixed memory from the temporally central latent frame of its first occurrence, avoiding updates from later clips that may accumulate drift. We aggregate each entity's cross-attention maps and binarize their top-k positions to obtain an entity mask. The scene mask is defined as the complement of the union of all character masks. Applying these masks to the pre-RoPE self-attention keys and the corresponding values selects the associated states, which are stored as

Table 1: **Quantitative baseline results.** Bold/underlined values mark the best/second-best results.

Setting	Method	Text Alignment↑	Dynamic Degree↑	Motion Smoothness↑	Intra-Clip Consistency↑	Inter-Clip Consistency↑	Shot Cut Accuracy↑	Time (s/frame)↓	GPU Peak (GB)↓
Hard	Infinity-RoPE	**9.10**	**36.33**	99.06	89.75	80.97	91.31	**0.17**	**28.71**
	Echo-Forcing	8.76	27.42	99.15	87.88	81.08	95.00	0.22	51.00
	MemActive	8.84	30.75	**99.41**	**94.40**	**85.99**	**98.13**	0.22	30.48
Smooth	Infinity-RoPE	9.24	22.92	99.31	95.59	89.65	N/A	**0.18**	28.71
	Echo-Forcing	9.16	5.08	**99.46**	95.83	89.74	N/A	0.23	51.10
	SWIFT	9.60	9.33	99.31	95.23	91.00	N/A	0.22	32.87
	MemRoPE	**9.73**	**54.42**	98.93	93.62	83.17	N/A	0.24	**25.76**
	MemActive	9.53	19.33	99.34	**96.18**	**91.98**	N/A	0.20	30.51

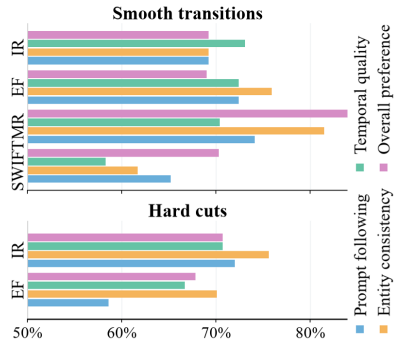


Figure 6: **User study results.**

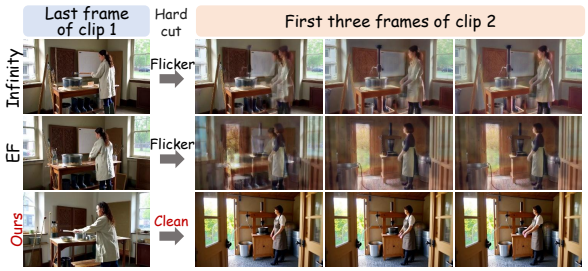


Figure 7: **MemActive enables full local-cache resets, yielding clean cut boundaries.**

$M_e = \left(\bar{K}_{m,e}, V_{m,e}\right)$. At inference time, the current structured prompt addresses the relevant entity memories, which RSQ-RoPE activates with entity-specific compression factors.

5 Experiments

5.1 Experimental Setup

Implementation details. We build MemActive on the Infinity-RoPE (Yesiltepe et al., 2026) inference pipeline using the Self Forcing (Huang et al., 2025) checkpoint based on Wan2.1-T2V-1.3B (Wan et al., 2025). We set $\lambda = 0.25$ for smooth transitions and $\lambda = 0.45$ for hard cuts. MemActive performs a full local-cache reset in hard cut mode while retaining separately stored SSM memories. Further details are provided in Appendix B.

Baselines. We primarily compare MemActive with training-free methods built on the same Self Forcing Huang et al. (2025). For smooth transitions, the baselines are Infinity-RoPE (IR) (Yesiltepe et al., 2026), Echo-Forcing (EF) (Wu et al., 2026), SWIFT (Tan et al., 2026), and MemRoPE (MR) (Kim et al., 2026). For hard cuts, we use IR and EF, which support this setting and are evaluated with their respective boundary-handling policies. Details are provided in Appendix F.

Evaluation data and metrics. We evaluate smooth-transition and hard-cut sets, each containing 100 prompt sequences of twelve 5-second clips spanning single- and multi-character and character/scene recurrence cases. We adapt VBench (Huang et al., 2024) to evaluate per-clip text alignment, dynamics, motion smoothness, and intra-clip consistency. Inter-clip Consistency measures recurring-appearance consistency using matched person crops and background patches, providing a quantitative measure of visual recall. Shot-Cut Accuracy (SCA) (Meng et al., 2026) measures boundary control under hard cuts. We also report inference time and peak GPU memory. All methods share prompts, output lengths, and random seeds. Details are in Appendix E.2.

5.2 Comparisons with Baselines

Quantitative comparison. Table 1 shows that MemActive achieves the best Inter-clip and Intra-clip Consistency in both transition modes, together with the best SCA under hard cuts. Figure 9 shows that MemActive's consistency decays slowest as the clip-index gap increases. Although IR attains a relatively high hard-cut Inter-clip score, Figure 8(a) shows that it often fails to switch to the prompt-specified scene or character.



Figure 8: **Qualitative comparison.** MemActive shows less visual drift when entities reappear across distant clips in both transition modes. Matching border colors denote the same character and scene.

Meanwhile, MemActive maintains competitive Text Alignment, Dynamic Degree, and Motion Smoothness, demonstrating a favorable overall balance. Although MemRoPE achieves the highest Dynamic Degree under smooth transitions, it has the lowest Inter-clip Consistency, illustrating the consistency–dynamics trade-off.

Qualitative comparison. Under hard cuts, separately stored entity memory enables a full local-cache reset, avoiding the flicker observed in IR and EF while preserving long-term recall (Figure 7). Appendix F details the boundary policies. Across nonadjacent clips, MemActive maintains character and scene consistency, whereas the baselines exhibit drift in both transition modes (Figure 8).

Human evaluation. We conduct a blinded pairwise study with 30 participants. Each participant evaluates 20 video pairs on four aspects, yielding 600 pairwise comparisons and 2,400 aspect-level judgments in total. The horizontal bars in Figure 6 report the percentage of judgments favoring MemActive, which exceeds 50% against every baseline across all four criteria and both transition modes. Protocol details are in Appendix E.3.

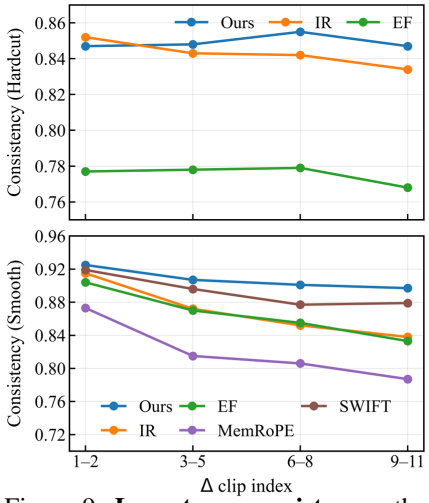


Figure 9: **Long-term consistency:** the Δ clip-index gap denotes the distance between each recurrence and the entity's first occurrence.

5.3 Training-Free and Training-Based Generalization

For training-free generalization, we apply IR (Yesiltepe et al., 2026) to Causal Forcing (CF) (Zhu et al., 2026) and AnyFlow (Gu et al., 2026) to obtain their Long Extension (LE) variants, then augment both with MemActive. For training-based generalization, we integrate MemActive into LongLive (Yang et al., 2026a) for smooth transitions and LongLive-2.0 (Chen et al., 2026) for hard cuts. MemActive improves Inter-clip Consistency across all four backbones (Table 2). MemActive

Table 2: **Generalization to other models.** Gray rows denote MemActive-augmented configurations. Bold marks the better result within each pair.

Setting	Model	Method	Text Alignment↑	Dynamic Degree↑	Motion Smoothness↑	Intra-Clip Consistency↑	Inter-Clip Consistency↑	Shot Cut Accuracy↑	Time (s/frame)↓	GPU Peak (GB)↓
					Training-free					
Hard	CF	Long Extension	**8.49**	95.33	**96.97**	85.46	78.53	88.52	**0.18**	**28.71**
		+ MemActive	7.77	**96.92**	96.85	**90.04**	**80.40**	**99.76**	0.22	30.48
	AnyFlow	Long Extension	7.58	10.25	**98.93**	**93.59**	82.01	90.24	**2.45**	**70.75**
		+ MemActive	**7.79**	**17.92**	98.91	92.91	**83.01**	**97.01**	2.76	75.97
Smooth	CF	Long Extension	**9.49**	95.25	**96.97**	85.48	78.77	N/A	**0.18**	**28.71**
		+ MemActive	8.43	**97.00**	96.87	**89.96**	**80.26**	N/A	0.20	30.51
	AnyFlow	Long Extension	8.47	39.92	99.00	**94.52**	83.16	N/A	**2.58**	**67.68**
		+ MemActive	**8.81**	**43.58**	**99.14**	94.48	**83.85**	N/A	3.13	72.95
					Training-based					
Hard	ShotStream	Original	6.04	2.75	99.60	96.47	87.89	77.09	0.28	48.77
	LongLive-2.0	Original	9.67	56.83	**99.41**	**92.36**	84.80	62.24	**0.87**	**75.03**
		+ MemActive	**10.06**	**61.50**	99.38	91.53	**86.18**	**72.15**	1.95	77.54
Smooth	AF	Original	8.93	4.08	99.43	95.67	91.96	N/A	0.26	32.85
	LongLive	Original	9.3102	7.2500	**99.3841**	**95.4942**	92.3051	N/A	**0.21**	**32.92**
		+ MemActive	**9.4470**	**7.3333**	99.3006	95.1171	**93.0850**	N/A	0.22	38.14

Table 3: **Quantitative ablation results.** Gray rows denote final configurations. Bold/underlined values mark the best/second-best results.

Setting	Variant	Fixed α	λ	Text Alignment↑	Dynamic Degree↑	Motion Smoothness↑	Intra-Clip Consistency↑	Inter-Clip Consistency↑	Shot Cut Accuracy↑	Time (s/frame)↓	GPU Peak (GB)↓
Hard	M_{base}	–	–	9.14	**43.92**	99.37	94.02	78.67	**100.00**	**0.20**	**29.88**
	M_{SSM}	1.00	–	**9.30**	38.75	99.39	93.99	79.24	99.52	0.21	30.48
	M^*_{full}	0.65	–	9.12	27.17	99.39	94.45	82.07	97.45	0.21	30.48
	M^*_{full}	0.45	–	9.11	25.83	99.41	94.50	83.39	95.44	0.21	30.48
	M^*_{full}	0.25	–	9.19	18.67	**99.45**	**95.10**	85.40	93.55	0.21	30.48
	M_{full}	–	0.65	8.80	29.58	99.40	94.48	82.38	98.17	0.22	30.48
	M_{full}	–	**0.45**	8.84	30.75	99.41	94.40	**85.99**	98.13	0.22	30.48
	M_{full}	–	0.25	8.79	23.50	99.42	94.73	84.54	95.79	0.22	30.48
Smooth	M_{base}	–	–	9.52	**22.58**	99.31	95.69	88.77	N/A	**0.18**	**28.99**
	M_{SSM}	1.00	–	9.52	15.67	99.39	96.11	90.37	N/A	0.21	30.50
	M^*_{full}	0.65	–	9.47	10.00	**99.44**	95.98	88.38	N/A	0.21	30.51
	M^*_{full}	0.45	–	9.41	13.25	99.43	95.91	88.81	N/A	0.21	30.51
	M^*_{full}	0.25	–	9.45	13.92	99.40	96.16	90.62	N/A	0.21	30.51
	M_{full}	–	0.65	9.51	15.42	99.37	96.04	90.25	N/A	0.20	30.51
	M_{full}	–	0.45	9.49	17.50	99.37	95.97	90.76	N/A	0.20	30.51
	M_{full}	–	**0.25**	**9.53**	19.33	99.34	**96.18**	**91.98**	N/A	0.20	30.51

surpasses Anchor Forcing (AF) (Yang et al., 2026b) in Inter-clip Consistency under smooth transitions and provides a more favorable balance of consistency, dynamics, and text alignment than ShotStream (Luo et al., 2026) under hard cuts. See Appendix H for more results.

5.4 Ablation Study

We denote the backbone, its SSM-only variant, and the full adaptive MemActive configuration by M_{base}, M_{SSM}, and M_{full}, respectively. For RSQ-RoPE, we evaluate fixed-α variants M^*_{full} with $\alpha \in \{0.25, 0.45, 0.65\}$ and adaptive variants M_{full} with $\lambda \in \{0.25, 0.45, 0.65\}$. All ablation variants use MemActive's full local-cache reset in hard-cut mode. We additionally compare RSQ-RoPE with direct attention-logit modulation using attention bias. Further details and results are provided in Appendix C.3 and Appendix J.

Effect of SSM. While adding SSM raises Inter-clip Consistency (*e.g.* from 78.67 to 79.24 in Table 3) under hard cuts, SSM alone remains insufficient. As shown in Figure 10, the recalled entity remains visually inconsistent in (a), while the red-boxed subject exhibits substantial appearance drift in (b).

Effect of RSQ-RoPE. Adding RSQ-RoPE after SSM substantially improves Inter-clip Consistency, confirming that correct memory retrieval alone does not ensure effective memory utilization. Each fixed-α variant uses a uniform compression strength, making it difficult to achieve a favorable balance between memory activation and other aspects of generation quality. In contrast, adaptive RSQ-RoPE relaxes this constraint, allowing the influence of retrieved memory to vary with its relevance. Empirically, the selected adaptive configurations achieve the best Inter-clip Consistency in both settings while providing a better overall balance across quality metrics.

Additive Bias as an Alternative to RoPE Remapping. Temporal RoPE compression improves recall alongside increased memory attention (Section 3). We therefore test whether directly boost-



Figure 10: **Qualitative ablation results.** In both modes, SSM improves consistency, with further gains from RSQ-RoPE.

Table 4: **Comparison with Attention Bias.**

Setting	Memory Activation	γ	Text Alignment↑	Dynamic Degree↑	Motion Smoothness↑	Intra-Clip Consistency↑	Inter-Clip Consistency↑	SCA↑
Hard cut	Attention Bias	1	8.57	28.17	99.40	94.34	82.81	95.04
		2	8.41	28.58	99.37	93.83	84.35	87.83
		4	7.72	18.00	99.26	90.97	83.84	80.21
	RSQ-RoPE (Ours)	–	**8.84**	**30.75**	**99.41**	**94.40**	**85.99**	**98.13**
Smooth	Attention Bias	1	9.39	13.67	**99.44**	96.04	89.50	N/A
		2	9.43	12.50	99.42	95.64	89.02	N/A
		4	**9.57**	**22.50**	99.20	93.40	84.89	N/A
	RSQ-RoPE (Ours)	–	9.53	19.33	99.34	**96.18**	**91.98**	N/A

ing memory logits can replace positional remapping. Keeping other settings unchanged, we add $b_e = \gamma s_e$ to each logit in entity e's memory group. In Table 4, RSQ-RoPE achieves higher Inter-clip Consistency than all tested bias settings in both modes. As γ increases, hard-cut consistency first rises and then falls, while smooth-transition consistency decreases. Unlike an entity-wise constant bias, RSQ-RoPE can alter the relative attention among memory tokens, a distinction further discussed and derived in Appendix C.3 alongside the bias formulation and experimental setup.

6 Limitations and Future Work

MemActive improves long-term consistency in prompt-interactive streaming video generation via selective entity-memory activation. However, maintaining multiple memories incurs inference overhead, while stronger activation may reduce motion dynamics and responsiveness to new prompts. Future work will explore sparse memory access and hierarchical RoPE-band modulation to further balance long-term recall with local temporal dynamics.

7 Conclusion

Temporal RoPE compression improves recall with fixed memory, accompanied by increased attention mass. MemActive combines SSM with RSQ-RoPE to adjust memory-route query positions while preserving native local positions. Experiments show improved consistency under hard cuts and smooth transitions, with gains extending to four additional backbones. These results identify positional compatibility as a factor in effective long-term memory utilization.

AI USE STATEMENT

Generative AI tools assist with drafting and editing portions of the manuscript and improving writing clarity. They are also used for literature search and discovery. In addition, they provide support for brainstorming and feedback on research ideas, methodology, and experimental design. We review and verify all methodological decisions, experimental results, claims, references, and AI-assisted content, and take full responsibility for the final content of this work.

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.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NeurIPS*, 2024.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. Anyflow: Any-step video diffusion model with on-policy flow map distillation. In *ECCV*, 2026.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *CVPR*, 2025.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. Longlive-rag: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *NeurIPS*, 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 *CVPR*, 2024.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Youngrae Kim, Qixin Hu, C-C Jay Kuo, and Peter Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *ECCV*, 2026.

Chen Li, Peng Zhang, Hanyu Zhou, Jialong Zuo, Fei Wang, Daiguo Zhou, Nong Sang, and Changxin Gao. Tether the subject, release the scene: Query-aware memory routing for long-horizon autoregressive video generation. *arXiv preprint arXiv:2608.26902*, 2026.

Jinzhuo Liu, Jiangning Zhang, Wencan Jiang, Yabiao Wang, Dingkang Liang, Zhucun Xue, Ran Yi, and Yong Liu. Advancing narrative long video generation via training-free identity-aware memory. *arXiv preprint arXiv:2605.18733*, 2026a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *ICLR*, 2026b.

Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling. In *ECCV*, 2026.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. In *CVPR*, 2026.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *ICLR*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Shanwen Tan, Hao Li, Jingtao Zhang, Xiaosong Jia, Xue Yang, Shaofeng Zhang, and Yanyong Zhang. Swift: Prompt-adaptive memory for efficient interactive long video generation. *arXiv preprint arXiv:2605.09442*, 2026.

Jiahao Tian, Yiwei Wang, Gang Yu, and Chi Zhang. Head forcing: Long autoregressive video generation via head heterogeneity. *arXiv preprint arXiv:2605.14487*, 2026.

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.

Mingqiang Wu, Weilun Feng, Zhefeng Zhang, Haotong Qin, Yuqi Li, Guoxin Fan, Xiaokun Liu, Zhulin An, Libo Huang, Yongjun Xu, et al. Echo-forcing: A scene memory framework for interactive long video generation. *arXiv preprint arXiv:2605.16003*, 2026.

Haiyang Xu, Zheng Ding, and Zhuowen Tu. Recap-forcing: Retaining content appearances for long video generation. *arXiv preprint arXiv:2608.26671*, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, YINGCONG CHEN, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *ICLR*, 2026a.

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region rope for interactive streaming video diffusion. 2026b.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *CVPR*, 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. 2026.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed KV-memory for consistent long video generation. In *BMVC*, 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. 2026.

APPENDIX

A SUMMARY

This appendix provides implementation details, theoretical analysis, evaluation protocols, and additional results. The contents are organized as follows:

- In Section B, we describe the implementation setup, memory construction and retrieval, and the main hyperparameter settings.
- In Section C, we analyze how temporal remapping changes the relative RoPE phase and explain why RSQ-RoPE is implemented on the query side.
- In Section D, we present the protocol and measurements used in the motivation experiments.
- In Section E, we describe the prompt suites, multi-shot evaluation metrics and aggregation, and human-evaluation protocol.
- In Section F, we provide implementation details for the compared baselines and clarify the adaptations used for fair comparison.
- In Section H, we report additional generalization results on training-free and training-based streaming video generators.
- In Section G, we compare MemActive with additional memory methods.
- In Section I, we illustrate generation with external reference images as memory.
- In Section J, we provide additional ablation results on the adaptive lower bound of RSQ-RoPE.
- In Section K, we discuss how a natural-language interface converts free-form instructions into the structured prompts used by MemActive.

B IMPLEMENTATION DETAILS

Generation Setup. We implement MemActive at inference time on the Infinity-RoPE (Yesiltepe et al., 2026) pipeline with the released Self Forcing (Huang et al., 2025) checkpoint based on Wan2.1-T2V-1.3B (Wan et al., 2025). Each clip is five seconds at 480×832 resolution and 16 FPS, generated with four denoising steps. Unless stated otherwise, we follow the original Infinity-RoPE generation configuration.

SSM Configuration. SSM constructs a fixed canonical memory for each character and scene from the temporally central latent frame of its first occurrence and does not refresh it with later generated clips. Entity masks are computed from cross-attention maps in layers 3–6. We max-normalize each selected-layer map, average the resulting maps across layers, and max-normalize the averaged map again before top-k selection. For each target character, we suppress positions associated with other co-occurring characters that have stored memories by subtracting the element-wise maximum of their masks and clamping negative values to zero. We then retain the top-k positions. The scene mask is the complement of the union of the character masks. We store the selected pre-RoPE keys and values at all Transformer layers, retaining at most 256 tokens per character and 1024 tokens per scene.

RSQ-RoPE Configuration. Context compatibility s_e is the cosine similarity between CLIP ViT-B/32 features of the description fields of the same persistent entity, clipped to $[0, 1]$. This score controls the activation strength of the stored appearance state in Eq. 5. All retrieved entity-memory groups are reindexed to a shared temporal anchor t_m, regardless of the clips from which they were extracted, while their native spatial coordinates are preserved. We set λ to 0.25 for smooth transitions and 0.45 for hard cuts.

C Analysis of Temporal RoPE Compression

C.1 Relative Temporal Phase in Self-Attention

For a token at spatiotemporal position $p = (t, w, h)$, let $R(t, w, h)$ denote the RoPE rotation applied to its pre-RoPE query or key. For a query q_i and key k_j, the self-attention logit is

$$\ell_{ij} = \frac{[R(t_i, w_i, h_i)q_i]^\top [R(t_j, w_j, h_j)k_j]}{\sqrt{d}}, \qquad a_{ij} = \mathrm{Softmax}_j(\ell_{ij}). \tag{8}$$

The value vectors are not rotated. In the motivation experiment, varying α changes only the remapped temporal coordinate, while each token's spatial coordinates remain at their native values and are unchanged across α settings. We therefore isolate the temporal RoPE component, denoted by $R_T(t)$, and omit the unchanged spatial terms below. For a query at temporal position t and a memory key at position m, we have

$$\langle R_T(t)q, R_T(m)k\rangle = q^\top R_T(m-t)k, \tag{9}$$

so changing the relative temporal offset changes the memory attention logit without changing the stored key or value.

Temporal compression reduces the query–memory phase offset from $\omega_r\Delta$ to $\omega_r\alpha\Delta$ at each RoPE frequency. For a query and memory key that are already aligned in content space, this smaller phase offset can increase their attention logit by improving positional compatibility. This observation provides intuition rather than a general monotonicity guarantee, since the full attention response also depends on feature directions, frequency components, and competing keys. The overall increase in memory attention is therefore established empirically in Figure 2.

C.2 Rationale for Query-Side Remapping

For clarity, we consider one retrieved entity and omit the entity subscript. Query-side remapping assigns the query the temporal coordinate

$$\tilde{t}_c = t_m + \alpha(t_c - t_m). \tag{10}$$

The resulting relative temporal phase is

$$t_m - \tilde{t}_c = \alpha(t_m - t_c). \tag{11}$$

The same phase can be obtained by keeping the query at t_c and assigning the memory key the coordinate

$$\tilde{t}_m(t_c) = t_c + \alpha(t_m - t_c) = (1-\alpha)t_c + \alpha t_m. \tag{12}$$

Therefore, query-side and key-side remapping produce the same RoPE phase for an individual query–memory pair. However, a generated chunk contains multiple query positions. Because $\tilde{t}_m(t_c)$ depends on t_c, exact key-side remapping would require a different rotated version of the same memory key for every query position. The memory keys could therefore no longer be represented by one shared rotated key sequence. RSQ-RoPE avoids this problem by keeping the memory keys fixed at their original memory position and constructing an entity-specific query view for each retrieved memory group. The local route continues to use the native query position. We therefore apply remapping on the query side because it preserves a reusable memory-key representation while leaving the pretrained query–local-context geometry unchanged.

C.3 Comparison with Attention Bias

For each retrieved entity e, the Attention Bias baseline uses the same relevance score s_e as RSQ-RoPE. Let $L_{m,e}$ denote the vector of pre-softmax logits for the keys in entity e's memory group. We modulate these logits as

$$\widetilde{L}_{m,e} = L_{m,e} + b_e\mathbf{1}, \qquad b_e = \gamma s_e, \tag{13}$$

where γ is a global coefficient controlling the bias strength and $\mathbf{1}$ broadcasts the scalar bias to every key in the group. We evaluate $\gamma \in \{1, 2, 4\}$. After concatenating the modulated entity groups into $\widetilde{L}_m$, we jointly normalize them with the unchanged local logits L_w:

$$[A_w \| A_m] = \mathrm{Softmax}([L_w \| \widetilde{L}_m]), \tag{14}$$

where A_w and A_m are the resulting local and memory attention weights. Queries and keys retain their native RoPE positions. Both methods use the same backbone and checkpoint, prompts, random seeds, memory construction and token selection, relevance scores, boundary policies, and generation and evaluation settings. They differ only in the memory-activation mechanism.

Effect on within-group attention. Consider a fixed query i and two memory keys j, k in the same entity group e, with pre-RoPE features held fixed. Let a_{ij} and a_{ij}^{bias} denote attention weights before and after adding b_e. Shared softmax normalization cancels in their ratio:

$$\frac{a_{ij}^{\text{bias}}}{a_{ik}^{\text{bias}}} = \frac{\exp(\ell_{ij} + b_e)}{\exp(\ell_{ik} + b_e)} = \exp(\ell_{ij} - \ell_{ik}) = \frac{a_{ij}}{a_{ik}}. \tag{15}$$

Thus, group-wise bias changes the group's weight relative to other context while preserving its conditional attention distribution.

For RSQ-RoPE, let $q_{i,T}$ and $k_{j,T}$ denote the temporal components of the pre-RoPE features, and let $\Delta_i = t_m - t_i$. All keys in group e share the temporal anchor t_m. Remapping the query to $\tilde{t}_{i,e} = t_m + \alpha_e(t_i - t_m)$ changes the relative temporal offset to $\alpha_e \Delta_i$. Since spatial contributions remain unchanged, the logit increment is

$$\delta\ell_{ij} = \ell_{ij}^{(\alpha_e)} - \ell_{ij} = \frac{q_{i,T}^\top D_{i,e} k_{j,T}}{\sqrt{d}}, \qquad D_{i,e} = R_T(\alpha_e \Delta_i) - R_T(\Delta_i). \tag{16}$$

Although $D_{i,e}$ is shared across the group, different key features can yield different logit increments. The resulting change in the attention ratio is

$$\log \frac{a_{ij}^{(\alpha_e)}}{a_{ik}^{(\alpha_e)}} - \log \frac{a_{ij}}{a_{ik}} = \frac{q_{i,T}^\top D_{i,e}(k_{j,T} - k_{k,T})}{\sqrt{d}}. \tag{17}$$

A nonzero right-hand side changes the within-group distribution even when the memory keys share a temporal index. This establishes a difference between the operations, rather than a guarantee of better token matching. The comparison holds for fixed features at one attention operation. Either intervention can also alter features in subsequent layers or generation steps. Both mechanisms can change local attention mass through shared normalization while retaining native local positions.

Relation to the empirical results. The controlled experiment in Section 3 measures the joint changes in memory attention and anchor consistency under RoPE compression. Table 4 instead tests additive bias as an alternative activation mechanism in the full multi-clip setting. Hard-cut Inter-clip Consistency first rises and then falls as γ increases, while smooth-transition consistency decreases across the tested values. RSQ-RoPE achieves higher Inter-clip Consistency than all tested bias settings in both modes. Attention Bias with $\gamma = 4$ yields higher Text Alignment and Dynamic Degree under smooth transitions, but lower Intra-clip and Inter-clip Consistency. These results support the effectiveness of RSQ-RoPE relative to the tested bias configurations. Isolating the contribution of within-group reweighting would require a comparison that controls memory attention mass.

D Implementation Details of Motivation Experiments

Protocol. We use a controlled two-clip experiment to isolate the effect of query–memory temporal remapping. We evaluate 100 cases in which the two clips share the same character and scene but use different prompts. At the clip boundary, we clear the cross-clip local KV states and use the temporally central latent frame of the first clip as the only cross-clip memory when generating the first three chunks of the second clip. We test $\alpha \in \{0.25, 0.50, 0.75, 1.00\}$ with the same random seed across all α settings within each case, so the settings differ only in their query–memory temporal mapping.

Measurements. We assess visual recall using consistency with the supplied anchor and characterize memory access using attention mass. Panel (a) of Figure 2 reports the attention allocation of each of the first three generated chunks under native RoPE ($\alpha = 1$), averaged over the 100 cases. Panel (b) reports memory attention and consistency with the memory anchor across the tested α values. The curves report the mean over the 100 cases, and the shaded regions denote the 25th–75th percentiles across cases. Memory consistency is the equal-weight average of the VBench Subject Consistency and Background Consistency scores (Huang et al., 2024), computed between each generated chunk and the memory anchor.



Figure 11: **MemActive generalizes across training-free streaming backbones.** Adding MemActive to the Long Extension (LE) configurations of Causal Forcing (Zhu et al., 2026) and AnyFlow (Gu et al., 2026) preserves recurring appearance across both hard-cut and smooth-transition sequences without changing the underlying backbone.

E Benchmark Details

E.1 Prompt Suites

We evaluate two suites, each containing 100 sequences of 12 five-second clips. Both suites cover single- and multi-character cases across diverse scenes. The hard-cut suite varies character–scene compositions across explicit scene changes and revisits recurring characters and scenes. The smooth-transition suite contains continuous prompt changes with single or multiple characters, camera-aware scenes, and changing actions.

E.2 Metrics and Aggregation

We evaluate all sequences with a multi-shot adaptation of VBench (Huang et al., 2024)[1]. We apply the original VBench implementations independently to each five-second clip and report all scores on a 0–100 scale.

Per-clip VBench Metrics. Text Alignment corresponds to VBench Overall Consistency and measures ViCLIP similarity between each clip and its prompt. Dynamic Degree reports the percentage of clips classified as non-static from RAFT optical flow, while Motion Smoothness measures reconstruction accuracy under AMT frame interpolation. We use the original VBench implementations without modification.

[1] https://github.com/Vchitect/VBench

Intra-clip Consistency. We use the official VBench Subject Consistency and Background Consistency metrics, which encode full frames with DINO ViT-B/16 and CLIP ViT-B/32, respectively.

Inter-character Consistency. We sample the middle frame of each clip and group clips by the character labels in their prompts. Faster R-CNN detects COCO person boxes, which are expanded by a factor of 1.2 before cropping. DINO-based exclusive matching associates person crops across clips within each character group while penalizing detection reuse and similarity to previously assigned tracks. We encode the matched crops with DINO ViT-B/16 and fall back to the full frame when no person is detected.

Inter-background Consistency. We group clips by scene label and divide each sampled frame into a 3×3 grid. We select up to four patches with the lowest person overlap, requiring each selected patch to overlap any detected person box by at most 5% of its area. We average their CLIP ViT-B/32 features to represent the background and fall back to the full frame when no eligible patch remains. For both inter-clip metrics, we average cosine similarities, clipped below at zero, over all cross-clip pairs that share the corresponding character or scene label.

Shot-Cut Accuracy. For hard cuts, we follow the SCA metric introduced by HoloCine (Meng et al., 2026). We use TransNetV2[2] to detect shot boundaries and match them one-to-one with the prompt-specified boundaries. We report $\mathrm{SCA} = 100\exp(-\mathrm{NSD})$, where $\mathrm{NSD} = (E_{\mathrm{matched}} + E_{\mathrm{penalty}})/F_{\mathrm{total}}$ measures matched-boundary displacement and penalizes missed or extra cuts relative to the total number of frames.

Aggregation. For each case, we average Text Alignment and Motion Smoothness over its 12 clips, and compute Dynamic Degree as the percentage of those clips classified as dynamic. Intra-clip Consistency is the equal-weight average of Subject Consistency and Background Consistency, while Inter-clip Consistency is the equal-weight average of Inter-character Consistency and Inter-background Consistency. We report the unweighted mean of the case-level scores over the 100 cases in each suite.

E.3 Human Evaluation Protocol

We conduct a blinded pairwise study with 30 participants. The pool contains 100 trials, and each participant evaluates 20 randomly sampled trials, yielding 600 pairwise comparisons. Each trial compares MemActive with one baseline under the same prompt. Method identities are hidden, and the A/B placement is randomized. Participants use a two-alternative forced-choice protocol to select the preferred video for prompt following, entity consistency, temporal quality, and overall preference. For recall trials, an earlier reference frame is shown with the later target clips, and entity consistency is judged for the character, scene, or character–scene pair recalled by the prompt. We report the percentage of judgments favoring MemActive for each setting, baseline, and aspect.

F Details of Comparisons with Baselines

We compare MemActive with Infinity-RoPE (Yesiltepe et al., 2026)[3], Echo-Forcing (Wu et al., 2026)[4], SWIFT (Tan et al., 2026)[5] and MemRoPE (Kim et al., 2026)[6] on the same Self Forcing (Huang et al., 2025) checkpoint. We evaluate Infinity-RoPE (Yesiltepe et al., 2026) and Echo-Forcing (Wu et al., 2026) under both hard-cut and smooth-transition protocols. SWIFT (Tan et al., 2026) and MemRoPE (Kim et al., 2026) support only smooth transitions and are evaluated under that protocol. Causal Forcing (Zhu et al., 2026) and AnyFlow (Gu et al., 2026) are reserved for the cross-model compatibility study in Section 5.3.

For hard-cut generation, Infinity-RoPE combines KV Flush, which retains the sink and recent frames to preserve short-term continuity while removing stale context, with RoPE Cut, which introduces a discontinuity in temporal RoPE coordinates to start a new scene (Yesiltepe et al., 2026). Echo-Forcing follows the same boundary-handling policy as Infinity-RoPE in our evaluation. The de-

[2] https://github.com/soCzech/TransNetV2

[3] https://github.com/yesiltepe-hidir/infinity-rope

[4] https://github.com/mingqiangWu/Echo-Forcing

[5] https://github.com/ShanwenTan/SWIFT

[6] https://github.com/YoungRaeKimm/MemRoPE



Figure 12: **MemActive generalizes to training-based streaming systems.** The qualitative comparisons cover hard cuts (top) and smooth transitions (bottom). Augmented configurations improve the preservation of recurring appearance.

scription above therefore also applies to Echo-Forcing. This coupled change in cached context and temporal coordinates produces transient flicker at the boundary. MemActive uses persistent SSM memory that is independent of the local and sink KV states, so it can fully clear these states without KV Flush or RoPE Cut while still retrieving long-term entity information. This decoupling enables a clean transition. We evaluate the baselines with their native policies to preserve their intended inference procedures.

G Additional Comparison with Memory Methods

Memory-augmented video generation methods can be analyzed along three complementary stages of the memory pipeline: memory maintenance and retention, memory selection and routing, and post-retrieval memory utilization. These stages are not mutually exclusive, and a method may operate at more than one stage.

(1) Memory maintenance and retention. This line progresses from fixed temporal cache layouts to adaptive and content-aware retention. LongLive (Yang et al., 2026a) combines a short local window with persistent frame sinks, while Infinity-RoPE (Yesiltepe et al., 2026) retains sink and recent states while reindexing their temporal positions. Deep Forcing (Yi et al., 2026) makes compression attention-aware, and Echo-Forcing (Wu et al., 2026) further separates scene memory into hierarchical components with difference-aware decay. Retention then becomes responsive to the evolving generation state: SWIFT (Tan et al., 2026) adapts its window and semantic anchors to prompt transitions, MemRoPE (Kim et al., 2026) continuously updates compressed memory tokens, and Anchor Forcing (Yang et al., 2026b) preserves anchor states for prompt-switch re-caching. Head Forcing (Tian et al., 2026) and Relax Forcing (Zhao et al., 2026) specialize retention by attention-head role and temporal function, respectively, whereas Echo-Infinity (Bian et al., 2026) replaces hand-designed updates with learned evolving memory queries. RECAP-Forcing (RF) (Xu et al., 2026) advances this direction by retaining states according to appearance novelty rather than temporal recency alone.

(2) Memory selection and routing. This line progresses from selecting relevant historical frames to semantic indexing and query-dependent access control. MemFlow (Ji et al., 2025) uses the current prompt to select historical frames and activate relevant memory tokens, while LongLive-RAG (Hu

Table 5: **Quantitative comparison with RECAP-Forcing.** Bold values indicate the better result within each transition setting.

Setting	Method	Text Alignment↑	Dynamic Degree↑	Motion Smoothness↑	Intra-Clip Consistency↑	Inter-Clip Consistency↑	Shot Cut Accuracy↑	Time (s/frame)↓	GPU Peak (GB)↓
Hard	RECAP-Forcing	7.78	20.67	99.02	91.87	**92.68**	58.67	0.24	**24.60**
	MemActive	**8.84**	**30.75**	**99.41**	**94.40**	85.99	**98.13**	**0.22**	30.48
Smooth	RECAP-Forcing	8.76	**20.00**	98.87	93.93	78.12	N/A	0.24	**24.59**
	MemActive	**9.53**	19.33	**99.34**	**96.18**	**91.98**	N/A	**0.20**	30.51

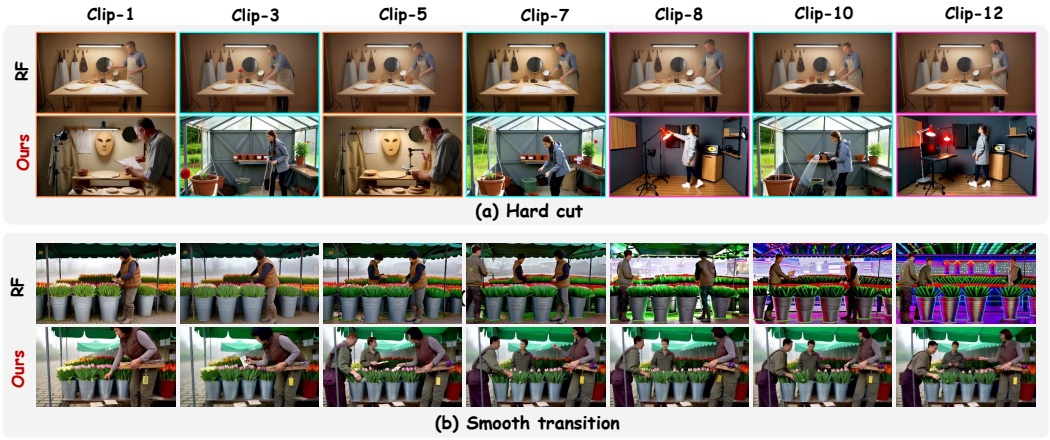


Figure 13: **Qualitative comparison with RECAP-Forcing.** Rows show RECAP-Forcing (RF) and MemActive. Under hard cuts (a), RF repeatedly preserves the initial scene instead of following the requested scene changes, whereas MemActive switches scenes at the requested boundaries while recalling recurring content. Under smooth transitions (b), RF degrades in later clips, whereas MemActive continues to generate visually coherent frames while preserving recurring-content consistency.

et al., 2026) treats generated latents as a searchable history for content-based retrieval. Entity-aware methods make these addresses explicit: IAMFlow (Liu et al., 2026a) tracks persistent entity identifiers, and MemActive's SSM uses structured character and scene identifiers to support direct retrieval and recombination. MemFlow (Ji et al., 2025) retrieves prompt-relevant historical frames and activates a selected subset of memory tokens, whereas TetherMem (Li et al., 2026) uses query-role-, region-, and age-conditioned attention priors to reweight access to historical tokens. These mechanisms improve memory access through selection or explicit attention routing.

(3) Post-retrieval memory utilization. Selection and routing determine which historical evidence a query can access or emphasize. MemActive studies a complementary factor that remains after relevant memory has been retrieved: whether the positional relation under RoPE allows a frozen generator to use that memory effectively. Given a fixed retrieved memory, RSQ-RoPE modifies the positional relation between each query and its entity-specific memory only along the memory route, while preserving the native query–local-context relation. Together with SSM, this design separates what memory is retrieved from how its positional compatibility with the current query is controlled.

Among these methods, we additionally compare with recent RF because it supports multi-prompt generation. Table 5 compares MemActive with RF under the same Self Forcing checkpoint. MemActive performs better on most quality metrics and achieves substantially higher Shot Cut Accuracy under hard cuts. Although RF obtains higher hard-cut Inter-clip Consistency (92.68 versus 85.99), its much lower SCA (58.67 versus 98.13) and Figure 13(a) show that it often retains the initial scene instead of following the requested cuts. Under smooth transitions, RF degrades in later clips, whereas MemActive maintains recurring content and achieves higher Intra-clip and Inter-clip Consistency (Figure 13(b)).



Figure 14: **External reference memory.**

H Additional Generalization Results across Generation Models

Figure 11 and Figure 12 show that MemActive generalizes across both training-free and training-based streaming backbones. The training-free results augment the Long Extension (LE) configurations of Causal Forcing (Zhu et al., 2026) and AnyFlow (Gu et al., 2026). The training-based results cover smooth transitions and hard cuts. Across these settings, MemActive preserves recurring character identities and scene appearance without modifying the original backbone checkpoints. These results demonstrate that MemActive generalizes across different generation backbones and training paradigms.

I External References as Memory

MemActive also stores external reference images as semantic memory using their pre-RoPE keys and values. Unlike the long-term setting, the memory source here is an external image rather than an earlier generated clip. The image is therefore addressed through the same memory interface, while the new prompt specifies the desired motion and scene evolution. This setting tests whether the memory route can preserve reference attributes when the requested action changes. Figure 14 illustrates that the generated clips retain the reference appearance while following the new motion prompts.

J Additional Ablation Study

Figure 15 qualitatively illustrates how the lower bound λ affects adaptive RSQ-RoPE. A smaller λ enables stronger recall of compatible entity memories. This better preserves recurring appearance, but may also carry stale appearance into the current clip. A larger λ favors adaptation to the current clip by limiting the influence of retrieved memories. However, it may weaken the recall of recurring appearance. This trade-off is observed under both hard cuts and smooth transitions. Based on this behavior, MemActive uses $\lambda = 0.45$ for hard cuts to reduce stale appearance and $\lambda = 0.25$ for smooth transitions to strengthen recurring-appearance recall.

K Natural-Language Interface

To improve user interaction in our engineering implementation, we provide a natural-language interface that allows users to describe each clip with free-form instructions instead of manually constructing structured prompts. An example is shown in Figure 16. The interface extracts the scene, active characters, their descriptions, and the requested action, and converts them into the structured prompt schema shown in Figure 5. It associates recurring characters and scenes with their persistent identifiers in the entity registry and assigns new identifiers to previously unseen entities. The resulting structured prompt is passed directly to SSM for memory construction and retrieval. This interface improves usability while leaving the underlying memory construction and RSQ-RoPE mechanisms unchanged.



Figure 15: **The adaptive lower bound λ controls the trade-off between recurring-appearance recall and adaptation to the current clip.**

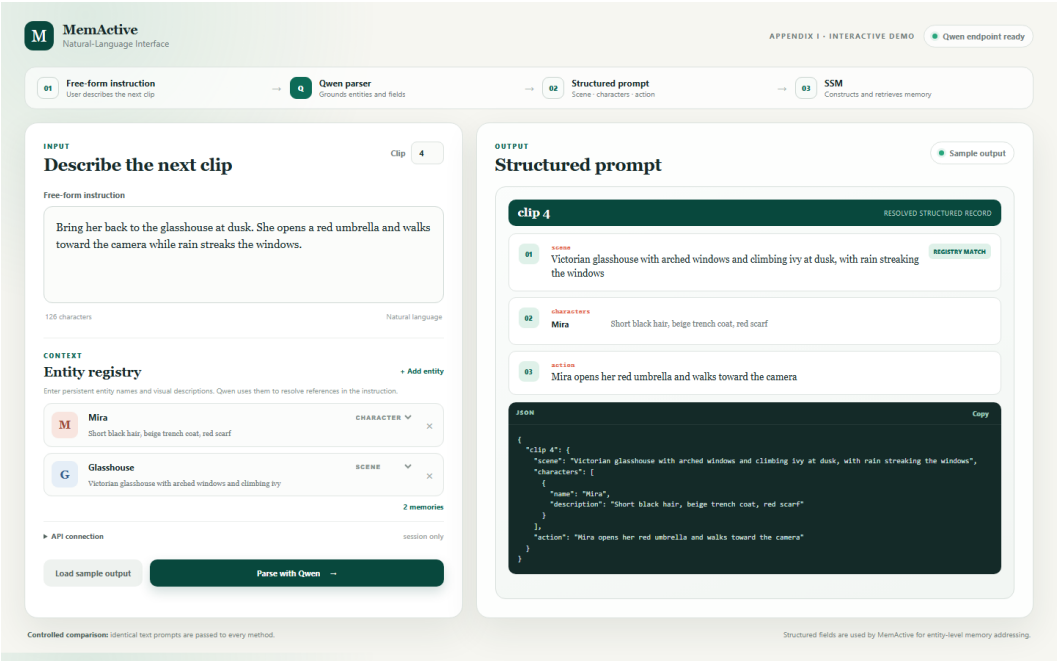


Figure 16: **Natural-language interface.** Users provide free-form instructions, which are automatically converted into structured prompts for memory construction and retrieval.