MEMFLOW: FLOWING ADAPTIVE MEMORY FOR CONSISTENT AND EFFICIENT LONG VIDEO NARRATIVES

Anonymous authors
Paper under double-blind review



Figure 1: Existing streaming interactive text-to-video models such as LongLive (Yang et al., 2025a) often fail to maintain consistency after prompt switching (suffering from redundant subjects or inter-clip inconsistency). MEMFLOW addresses this by maintaining dynamic memory for long-term consistency, enabling narrative coherence even in challenging cases which include new subjects appearance or scenario switches.

ABSTRACT

The core challenge for streaming video generation is maintaining content consistency over long context, which poses high requirement for the memory design. Most existing solutions maintain the memory by compressing historical frames with predefined strategies. However, different to-generate video chunks should refer to different historical cues, which is hard to satisfy with fixed strategies. In this work, we propose MEMFLOW to address this problem. Specifically, before generating the coming chunk, we dynamically retrieve the most relevant historical frames as context with the text prompt of this chunk from the global memory bank. This design not only accurately sources the context needed to maintain visual consistency, but also ensures semantic coherence even as new events unfold or scenes transition. In addition, during generation, we only activate the most relevant tokens in memory for each query in the attention layers, which effectively guarantees the generation efficiency. After tuning the model on curated memory-intensive scenarios to unlock its ability to leverage distant historical cues, MEMFLOW achieves outstanding long-context consistency with minimal computation overhead (7.9% speed reduction compared with the memory-free baseline) and is designed for compatibility with any streaming video generator using a KV cache.

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

Video generation has attained remarkable quality (Yang et al., 2024; Kuaishou, 2024; Kong et al., 2024; Wan et al., 2025; OpenAI, 2024), making its extension to long durations critical for advancing creative and cinematic applications. While Diffusion Transformer (DiT) models (Peebles & Xie, 2023) leverage bidirectional attention to capture complex spatiotemporal dependencies, their inherent computational costs and GPU memory limits constrain them to short video generation. Autoregressive (AR) diffusion models (Teng et al., 2025; Yin et al., 2025; Huang et al., 2025b; Chen et al., 2025b) offer a promising alternative by decomposing long videos into sequential clips, which alleviates the computation bottleneck through a reduced attention window.

Owing to their sequential nature, AR models are naturally suited for interactive video generation, which has emerged as a crucial application for enabling users to guide narratives with streaming prompts. Most existing works conduct chunk-level autoregressive generation—conditioning new segments on past content and current prompt—to seamlessly introduce new elements and scene switches over extended durations. However, this paradigm poses difficulties in effectively preserving and leveraging memory for long-range content consistency due to complex inter-clip dependencies. First, since different to-generate video chunks should refer to different historical cues, the memory is required to adaptively provide relevant context according to streaming prompts; Second, the capacity of stored memory must be highly constrained, a necessity dictated by both the hardware limits of GPU memory and the demands of generation efficiency. Last, the model must effectively learn how to utilize the context provided by the memory for consistent generation. Existing approaches fail to fully address these challenges. They preserve memory in predefined paradigms—employing the first chunk as memory sink (Yang et al., 2025a), storing historical frames through fixed compression (Zhang & Agrawala, 2025; Gu et al., 2025; Xiao et al., 2025a), or baking context implicitly with trainable modules (Jiang et al., 2025; Dalal et al., 2025; Hong et al., 2024)—yet these rigid strategies cannot dynamically fetch historical content tailored to varying prompts, and their training procedures fail to elicit the model's capacity to harness diverse memory.

Thus, we propose Narrative Adaptive Memory (NAM). Unlike rigid buffers, NAM maintains a global memory bank continuously updated by historical visual tokens (KV cache) from generated chunks, and dynamically retrieves the most semantically relevant context for the current text prompt—such as a specific character's identity when it reappears. NAM preserves narrative coherence as new events emerge or scenarios switch, a challenge for fixed memory strategies.

However, the introduction of memory inevitably brings extra computation burden, as video token counts are vastly larger than those in text or image tasks. Thus, we propose Sparse Memory Activation (SMA), which strategically activates only the most relevant tokens in attention layers according to the attention scores calculated from query (current chunk) and key (retrieved memory) by top-k selection. Attention over only these tokens accelerates inference while preserving quality.

Furthermore, zero-shot incorporation of distant historical tokens is insufficient—the model must be explicitly trained to unlock the memory utilization capability. To this end, we introduce Memory-Forced Streaming Tuning (MST). MST employs curated memory-intensive prompts (e.g., scene re-visitation) during training to explicitly force the model to fetch distant yet crucial cues from the memory bank for consistent generation, thereby enabling dynamic memory use.

In this way, our MEMFLOW effectively maintains contextual consistency over long durations and adeptly balances the memory–efficiency trade-off. It achieves state-of-the-art quality for interactive video generation with only 7.9% speed reduction compared with memory-free baseline. Our framework sustains 18.7 FPS on a single NVIDIA H100, demonstrating clear advantages in narrative-coherent, long-term consistent videos with complex character and scene switching.

2 RELATED WORKS

Long Video Generation. Prior efforts can be broadly categorized into three approaches. Autoregressive-diffusion hybrids (Yin et al., 2025; Huang et al., 2025b; Chen et al., 2025b; Teng et al., 2025; Zhang & Agrawala, 2025; Yang et al., 2025a) generate long videos by iteratively predicting frames (Yin et al., 2025; Huang et al., 2025b; Chen et al., 2025b; Teng et al., 2025; Zhang & Agrawala, 2025; Yang et al., 2025a). Diffusion-Forcing (Chen et al., 2025a) mitigates error propagation by adjusting denoising schedules. CausVid (Yin et al., 2025) distills bidirectional models

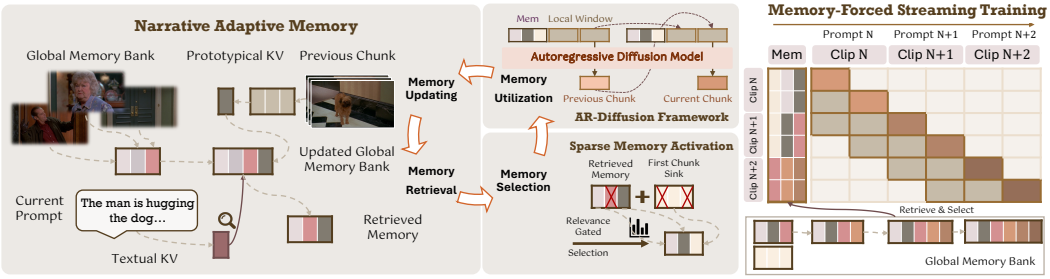

Figure 2: **Overall Framework of MEMFLOW,** which is designed for interactive long video generation with both long-term consistency and efficiency. In autoregressive generation, we conduct memory updating, retrieval, and selection sequentially before finally use it for synthesizing (Sec 3.1). After including the previous chunk to update global memory bank, we first use textual token from current prompt to query the memory for retrieving semantically aligned KV cache (Sec 3.2). The retrieved memory is then combined with "Frame Sink"-the KV cache from the first chunk-for the following Sparse Memory Activation, which filters the most relevant context for current chunk generation (Sec 3.3). Memory-Forced Streaming Tuning is designed for adequately stimulating the model's ability to exploit dynamic memory.

into few-step causal models and Self Forcing (Huang et al., 2025b) further addresses the train–test gap. Multistage methods (Xiang et al., 2025; Zhuang et al., 2024; Huang et al., 2025a; Xiao et al., 2025a) decompose long videos into separately generated clips via keyframe synthesis (Zhou et al., 2024; Huang et al., 2024a; Xiao et al., 2025a) or sequential prompting (Zhao et al., 2025; Long et al., 2024), but their isolated clip generation often lacks temporal coherence. The third category applies efficient architectures (Dalal et al., 2025; Zhang et al., 2025b; Ouyang et al., 2025) to manage computational costs via linear attention or condensed tokens, but often sacrifice visual fidelity.

Memory Mechanisms in Video Generation. Effective memory mechanisms are crucial for maintaining consistency in long video generation. Action-guided video generation often relies on geometric and spatial dependencies (Zhai et al., 2025; Chen et al., 2025c). Most of them (Sun et al., 2025; Wang et al., 2026) maintain memory banks and retrieve context based on camera-pose overlap (Xiao et al., 2025b; Yu et al., 2025), but are highly dependent on spatial priors and lack generalizability. General video generation primarily maintains memory through context compression (Zhang & Agrawala, 2025; Gu et al., 2025; Hong et al., 2024): FramePack (Zhang & Agrawala, 2025) and RELIC (Hong et al., 2025) compress input frames into fixed-size context to manage memory and efficiency, while FAR (Gu et al., 2025) and PFP (Zhang et al., 2025a) combine short- and long-term memory via multiscale compression and learnable modules, respectively. These methods maintain memory without adaptive retrieval, making it challenging to build dynamic connections between relevant context and the currently generated clip.

3 METHOD

MEMFLOW enhances long-video narrative consistency by incorporating a novel dynamic memory bank into a Memory-Forced Streaming Tuning framework (MST) (Sec 3.4). To dynamically recall relevant historical context according to current prompt, we first adopt (1) Narrative Adaptive Memory (NAM) mechanism for memory retrieval (Sec 3.2); then (2) Sparse Memory Activation (SMA) is employed for memory selection to address memory-efficiency trade-off (Sec 3.3). The refreshed memory is subsequently utilized by AR-diffusion model to synthesize the current video chunk. This process continues to roll out over extended temporal horizon in both training and inference, enabling the model to learn how to effectively manage and apply its memory over long durations. Figure 2 provides a high-level illustration of our framework (Sec 3.1).

3.1 OVERALL FRAMEWORK

Our work builds upon a hybrid autoregressive-diffusion framework that integrates autoregressive chunk-wise video generation with denoising diffusion (Yin et al., 2025; Teng et al., 2025; Huang et al., 2025b). At each generation iteration, the model produces a chunk of T frames, conditioned on the N immediately preceding frames. The causal attention allows Key-Value (KV) cache for

efficient inference, which we naturally leverage as the foundational structure of our memory bank without extra computational overhead. In a standard setup, attention operates over $(N + T)$ local frames. By integrating a memory of B frames, we extend attention to span $(B + N + T)$ frames, seamlessly blending short-term dependencies and long-term memory.

We follow Self Forcing (Huang et al., 2025b) to apply Distribution Matching Distillation (DMD) (Yin et al., 2024b) with a pretrained bidirectional teacher, and adopt LongLive (Yang et al., 2025a)'s streaming pipeline that conducts autoregressive rollout with streaming prompt inputs during optimization. Building upon this, we introduce two key memory components and a dedicated training strategy: (1) Narrative Adaptive Memory (NAM) retrieves semantically relevant frames from self-generated history, dynamically acquiring contextual memory for generation; (2) Sparse Memory Activation (SMA) guarantees computational efficiency by selectively activating only the most relevant memory tokens; (3) Memory-Forced Streaming Tuning (MST) employs curated memory-intensive prompts during distillation to explicitly force the model to fetch distant cues from the memory bank, enabling dynamic memory use for long-range consistency.

3.2 Narrative Adaptive Memory (NAM)

We first formalize the components of NAM. During the autoregressive rollout of generation, we maintain a global memory bank of Key-Value (KV) cache from synthesized frames $\{\boldsymbol{K}^l, \boldsymbol{V}^l\}_{l=1}^L$, for the l-th transformer layer of DiT. To sustain long-range context, this bank is refreshed over autoregressive rollouts. At iteration m, NAM dynamically retrieves a subset from $\{\boldsymbol{K}_m^l, \boldsymbol{V}_m^l\}_{l=1}^L$ as context for attention computation to ensure the context length remains within GPU capacity, To mitigate the rapid expansion of the global memory while preserving salient cues, we introduce two synergistic strategies: **(i) Periodic Update**, which refreshes the global memory in a highly sparse manner at a fixed temporal stride; and **(ii) Semantic Retrieval**, which extracts the most informative historical context based on semantic relevance. Together, these techniques establish long-term dependencies for narrative consistency within a fixed memory footprint.

Periodic Update. We posit that the memory required for long-term consistency primarily encapsulates key visual identities, such as the appearance of subjects and background. Given the high temporal redundancy across consecutive video chunks and within those short chunks, a single representative frame is sufficient to encapsulate the core visual attributes of the entire segment. Specifically, at prescribed intervals for updating, the KV pairs of the first frame from the generated chunk serve as a representative prototype for integrating into the global memory bank. The updated bank is formed by concatenating the existing historical frames with the newly selected prototype, maintaining computational efficiency without losing salient information.

Semantic Retrieval. To provide content-aligned context for generation, we employ a retrieval criterion derived from the cross-attention scores between the textual tokens of the current prompt (query) and the visual tokens in the global memory bank (key), which has proven effective in prior works (Chen et al., 2024; Wang et al., 2025; Yang et al., 2025b) of large vision-language models. We expect the visual tokens that yield higher attention scores to be more semantically aligned with the chunk currently being synthesized.

Let $\boldsymbol{Q}_{\text{text}}^l \in \mathbb{R}^{t\times d}$ be the textual query tokens of the current prompt at layer l, where t is the number of text tokens. For each of the b frames stored in the global bank, represented by its KV cache key $\boldsymbol{K}_{m,i}^l \in \mathbb{R}^{n\times d}$ (where n is the number of visual tokens per frame and $i = 1, \ldots, b$), we compute a semantic relevance score $\mathcal{S}_{m,i}^l$ as:

$$\begin{aligned} \mathcal{S}_{m,i}^l &= \text{Aggregate}\left(\boldsymbol{A}_{m,i}^l\right), \\ \boldsymbol{A}_{m,i}^l &= \text{Softmax}\left(\frac{\boldsymbol{Q}_{\text{text}}^l (\boldsymbol{K}_{m,i}^l)^\top}{\sqrt{d}}\right) \in \mathbb{R}^{t\times n}, \end{aligned} \tag{1}$$

where $\text{Softmax}(\cdot)$ normalizes each text token's scores over all visual tokens, and $\text{Aggregate}(\cdot)$ averages $\boldsymbol{A}_{m,i}^l$ over both dimensions to produce a scalar score $\mathcal{S}_{m,i}^l \in \mathbb{R}$ reflecting the semantic affinity between the prompt and the historical frame. We identify the indices of the top-K most semantically aligned frames:

$$\mathcal{I}_K = \underset{I\subseteq\{1,\ldots,b\},|I|=K}{\arg\max} \sum_{i\in I} \mathcal{S}_{m,i}^l, \tag{2}$$

to form the retrieved memory $\{\boldsymbol{K}_m^{'l}, \boldsymbol{V}_m^{'l}\} = \{\boldsymbol{K}_{m,i}^{l}, \boldsymbol{V}_{m,i}^{l}\}_{i\in\mathcal{I}_K}$. Note that NAM retrieves independently per layer, allowing each transformer layer to select its own most relevant historical frames.

3.3 Sparse Memory Activation (SMA)

Directly extending local context window to incorporate additional memory introduces a computational burden, as attention complexity scales with context size. While rigidly compressing the context can improve efficiency, it often compromises memory quality, as critical historical cues may be discarded indiscriminately. To address this memory–efficiency trade-off, we introduce Sparse Memory Activation, a relevance-gated technique for dynamic memory pruning before attention computation.

Our approach operates on the principle of selective attention: at each denoising step, the query visual token from the current video chunk attends only to a subset of the most relevant historical frames in memory. We compute a compact descriptor for both the query ($\boldsymbol{Q}_{\text{vis}}$) from the current chunk and the key of each frame from the NAM-retrieved memory $\{\boldsymbol{K}^{'}, \boldsymbol{V}^{'}\}$, using mean pooling over the token dimension, which is highly sufficient and expressive for generation tasks as demonstrated by prior works (Cai et al., 2025). The chunk index m and layer index l are omitted here for simplicity. This yields a single query descriptor, $\bar{\boldsymbol{q}}_{\text{vis}} \in \mathbb{R}^{1\times d}$, and a set of frame-wise key descriptors, $\{\bar{\boldsymbol{k}}_j\}_{j=1}^{b} \in \mathbb{R}^{1\times d}$. The relevance between the current query and frame-wise key in memory is then determined by the inner product of their respective descriptors:

$$r_j = \bar{\boldsymbol{q}}_{\text{vis}}^{\top}\bar{\boldsymbol{k}}_j, \quad \text{for} \quad j \in \mathcal{I}_K \tag{3}$$

Based on these relevance scores, we identify the set of indices $\mathcal{J}_k$ corresponding to the top-k most relevant frames:

$$\mathcal{J}_k = \underset{J\subseteq\mathcal{I}_K, |J|=k}{\arg\max} \sum_{j\in J} r_j \tag{4}$$

This formulation selects the subset of indices J of size k that maximizes the sum of relevance scores. Finally, the attention computation for query $\boldsymbol{Q}_{\text{vis}}$ is restricted to the key-value pairs belonging to the selected top-k frames:

$$\text{Attn}(\boldsymbol{Q}_{\text{vis}}, \boldsymbol{K}^{'}, \boldsymbol{V}^{'}) \approx \text{Attn}(\boldsymbol{Q}_{\text{vis}}, \boldsymbol{K}_{\mathcal{J}_k}^{'}, \boldsymbol{V}_{\mathcal{J}_k}^{'}) \tag{5}$$

where $\boldsymbol{K}_{\mathcal{J}_k}^{'}$ and $\boldsymbol{V}_{\mathcal{J}_k}^{'}$ are the key and value from the frames indexed by $\mathcal{J}_k$. SMA dynamically activates relevant tokens using the current denoising step's query, ensuring that each step attends to the most pertinent context. By selectively activating memory, SMA alleviates computational overhead and accelerates generation while preserving salient context. It balances the efficiency-memory trade-off without quality degradation.

3.4 Memory-Forced Streaming Tuning (MST)

While the introduced memory provides the architectural foundation for long-context generation, the model still needs to acquire the capability to actively utilize this memory as context. Recent work like LongLive (Yang et al., 2025a) implements a streaming long tuning procedure that learns on long videos while keeping supervision local. However, simply injecting a memory module into this paradigm does not guarantee that the student model will inherently learn to exploit it. Without explicit constraints, the model may still predominantly rely on the short-term local window for continuous generation while ignoring the long-term memory.

To explicitly elicit the model's memory-referencing capability, we propose a tailored streaming tuning strategy. Rather than using text sequences with only simple subject or camera movements as in LongLive (Yang et al., 2025a), we curate specific streaming prompts that deliberately depict scenarios requiring memory retrieval, such as character re-appearance or scene re-visitation. During training, these carefully designed prompts serve as strong conditions that force the student model to synthesize video chunks relying on previously stored memory. Under these memory-intensive cases, the DMD-based supervision effectively compels the student model to fetch distant yet crucial information from the memory bank to maintain long-range consistency, rather than falling back on the local window that primarily maintains short-term continuity.

Consequently, even if the local window lacks the critical visual cues necessary for generation (e.g., specific character identities or fine-grained environmental layouts that appeared several chunks

Table 1: **Quantitative comparison under multi-prompt 60-second setting** with representative long video generation models, where SkyReels-V2 (Chen et al., 2025b), Self Forcing (Huang et al., 2025b) and FramePack (Zhang & Agrawala, 2025) are adapted for the task by switching prompts.

Method	Quality Score ↑	Aesth. Drift ↓	Subject Consist. ↑		Scene Consist. ↑		Semantic Consist. ↑	
			CLIP	DINO	CLIP	DINO	Prompt	Transition
SkyReels-V2 (Chen et al., 2025b)	0.7771	0.0913	0.4137	0.3014	0.3780	0.2797	0.1765	0.6667
Self Forcing (Huang et al., 2025b)	0.7739	0.1084	0.4972	0.3703	0.3904	0.2796	0.1767	0.7102
LongLive (Yang et al., 2025a)	0.7480	0.0673	0.3147	0.2700	0.5052	0.4644	0.1803	0.1079
FramePack (Zhang & Agrawala, 2025)	0.7821	0.0399	0.3890	0.2943	0.4030	0.3021	**0.1868**	0.5059
MEMFLOW	**0.7849**	**0.0314**	**0.5689**	**0.4884**	**0.5150**	**0.4674**	0.1827	**0.7216**

Table 2: **Quantitative comparisons under single-prompt 5-second setting** with representative open-source video generation models of similar parameter sizes and resolutions. Our MEMFLOW performs on par with other models on overall quality and has a clear advantage in semantic score attributed to textual retrieval-based memory. † denotes the values reproduced by us.

Model	#Params	Resolution	Throughput (FPS) ↑	Evaluation scores ↑		
				Total	Quality	Semantic
Diffusion models						
LTX-Video (HaCohen et al., 2025)	1.9B	768×512	8.98	80.00	82.30	70.79
Wan2.1 (Wan et al., 2025)	1.3B	832×480	0.78	84.26	85.30	80.09
Autoregressive models						
SkyReels-V2 (Chen et al., 2025b)	1.3B	960×540	0.49	82.67	84.70	74.53
MAGI-1 (Teng et al., 2025)	4.5B	832×480	0.19	79.18	82.04	67.74
CausVid (Yin et al., 2025)	1.3B	832×480	17.0	81.20	84.05	69.80
NOVA (Deng et al., 2025)	0.6B	768×480	0.88	80.12	80.39	79.05
Pyramid Flow (Jin et al., 2025)	2B	640×384	6.7	81.72	84.74	69.62
Self Forcing, chunk-wise (Huang et al., 2025b)	1.3B	832×480	17.0	84.31	85.07	<u>81.28</u>
Self Forcing, frame-wise (Huang et al., 2025b)	1.3B	832×480	8.9	84.26	85.25	80.30
LongLive (Yang et al., 2025a)	1.3B	832×480	**20.3**†	<u>84.87</u>	**86.97**	76.47
MEMFLOW	1.3B	832×480	<u>18.7</u>	**85.14**	<u>85.95</u>	**81.90**

prior), the model resorts to the historical context to sustain narrative coherence. This memory-aware training mechanism effectively bridges the gap between architectural design and model capability, maximizing the utility of the proposed dynamic memory framework.

4 EXPERIMENT

4.1 IMPLEMENTATION DETAILS

We build MEMFLOW on Wan2.1-T2V-1.3B (Wan et al., 2025) equipped with adaptive memory and sparse activation. We initialize our model using the checkpoint from the first DMD stage of LongLive (Huang et al., 2025b) and conduct memory-forced streaming tuning with LoRA (Hu et al., 2022) of rank 256 for 3000 steps. During training, each iteration continues the model's own rollout by generating the next 5s video clip until 60s. Each batch includes five prompt switches with the switch time sampled uniformly from a predefined list from 5s to 55s, using a tailored prompt dataset constructed by Gemini-2.5-pro.

4.2 COMPARISONS FOR MULTI-PROMPT GENERATION

Since our approach is designed for interactive long video generation with multiple prompts, we first compare MEMFLOW's abilities with representative long video generation models. For fair comparison, we adapt existing methods - including SkyReels-V2 (Chen et al., 2025b), FramePack (Zhang & Agrawala, 2025), and Self Forcing (Huang et al., 2025b) - for multi-prompt video generation by switching prompts during autoregressive synthesis, where Self Forcing recomputes the KV cache at each prompt switch following LongLive (Yang et al., 2025a), which inherently supports generating videos with interactive instructions. We curate 100 sets of narrative scripts, each comprising six successive 10s prompts, resulting in 100 videos of 60 seconds each. Notably, every script is specifically designed to feature the disappearance and subsequent reappearance of both primary subjects and backgrounds, thereby providing a rigorous evaluation of the model's memory performance.

Table 3: **User study** under multi-prompt 60s setting. Values report human preference rate (%) for each method across three criteria: Visual Quality (VQ), Instruction Following (IF), and Global Consistency (GC).

Method	VQ	IF	GC
SkyReels-V2 (Chen et al., 2025b)	13.0	15.6	15.6
Self-Forcing (Huang et al., 2025b)	2.6	1.3	1.3
LongLive (Yang et al., 2025a)	23.4	23.4	22.1
FramePack (Zhang & Agrawala, 2025)	1.3	1.3	1.3
MEMFLOW	**59.7**	**58.4**	**59.7**

Table 4: **Quantitative comparisons for single-prompt 30-second setting** with representative long video generation models, showing more pronounced superiority than 5-second setting on all metrics with efficiency comparable to state-of-the-art models.

Model	Total Score ↑	Quality Score ↑	Semantic Score ↑	FPS ↑
SkyReels-V2 (Chen et al., 2025b)	75.29	80.77	53.37	0.49
FramePack (Zhang & Agrawala, 2025)	81.95	83.61	75.32	0.92
Self Forcing (Huang et al., 2025b)	81.59	83.82	72.70	17.0
LongLive (Yang et al., 2025a)	83.52	85.44	75.82	**20.3**
MEMFLOW	**84.51**	**85.92**	**78.87**	18.7



Figure 3: **Qualitative comparisons under multi-prompt 60-second setting** with representative long video generation models, where MEMFLOW outperforms other alternatives in narrative coherence and subject consistency, without drifting or incoherent transition.

We evaluate narrative video generation on overall quality using VBench-Long (Huang et al., 2024c) quality score and long-term consistency via three memory-centric metrics computed on 10-second video clips segmented by prompt boundaries: (1) *Visual Drift*: quality degradation is measured by aesthetic score difference between first and last segment (Schuhmann et al., 2022); (2) *Cross-segment Consistency*: CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) similarity of SAM-3 (Carion et al., 2025) segmented subjects/backgrounds across segments—for segments with shared identities, features are clustered for similarity computation, while a penalty is applied to transition segments where original entities erroneously persist; (3) *Semantic Consistency*: per-segment prompt alignment via ViCLIP (Wang et al., 2023) cosine similarity and cross-segment transition alignment verifying subject presence/absence per script. Details are provided in the Appendix.

Quantitative Results. Table 1 shows that MEMFLOW achieves the best overall quality among all methods, verifying its comprehensive competitiveness. It also achieves top scores across almost all memory-centric metrics: although Framepack (Zhang & Agrawala, 2025) performs slightly better on prompt alignment, MEMFLOW's significant advantage in transition alignment shows that our method precisely follows the script to generate coherent narratives with complex plots including subject injection and scenario switches. Meanwhile, cross-segment consistency demonstrates that our method outperforms all other models in maintaining high subject and scene consistency across segments, despite challenging cases of character reappearance and background shift, demonstrating the effectiveness of our memory mechanism in preserving crucial visual identities over long horizons. Our model also displays substantially lower aesthetic drift, highlighting its superiority in mitigating error accumulation during long rollouts.

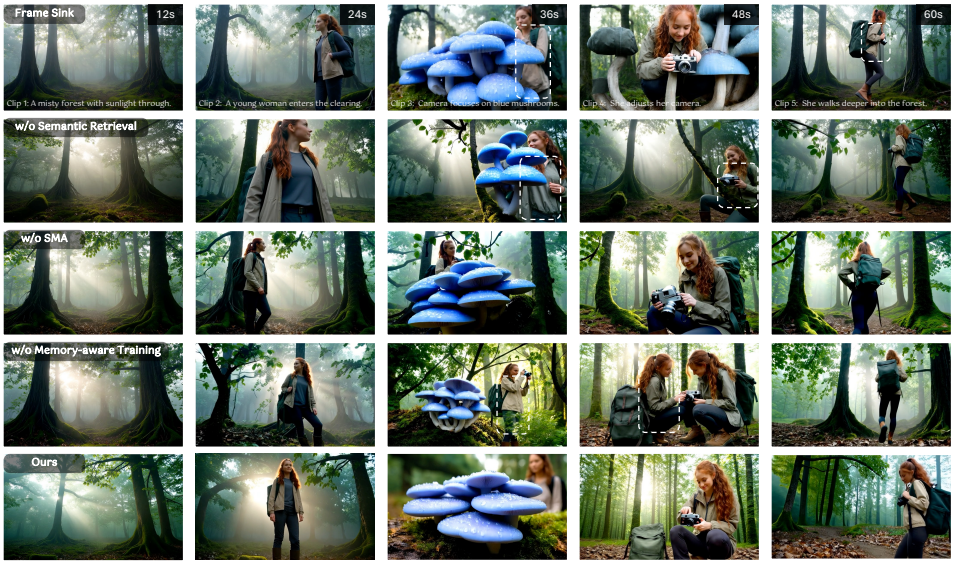

Figure 4: **Qualitative analysis of different memory mechanisms** under multi-prompt 60-second setting. "Frame Sink" retains KV cache from the first chunk as memory (Yang et al., 2025a); "w/o Semantic Retrieval" maintains the entire memory bank without retrieval; "w/o SMA" disables sparse activation; "w/o MST" removes memory-aware training; "MEMFLOW" is our full model.

Table 5: **Ablation study of memory mechanisms** under multi-prompt 60-second setting. "Frame Sink" retains KV cache from the first chunk as memory (Yang et al., 2025a); "w/o Semantic Retrieval" maintains the entire memory bank without retrieval; "w/o SMA" disables sparse activation; "w/o MST" removes memory-aware training; "MEMFLOW" is our full model.

Method	Quality Score ↑	Aesthetic. Drift ↓	FPS ↑	Subject Consist. ↑		Scene Consist. ↑		Semantic Consist. ↑	
				CLIP	DINO	CLIP	DINO	Prompt	Transition
Frame Sink (Yang et al., 2025a)	0.7480	0.0673	**20.3**	0.3147	0.2700	0.5052	0.4644	0.1803	0.1079
w/o Semantic Retrieval	0.7730	0.0453	13.1	0.5338	0.3767	0.4996	0.4355	0.1813	0.7045
w/o SMA	**0.7878**	0.0337	17.6	0.5613	0.4727	**0.5221**	**0.4684**	**0.1843**	0.7018
w/o MST	0.7700	0.0485	18.7	0.5684	0.4185	0.4691	0.3804	0.1810	0.6818
MEMFLOW	0.7849	**0.0314**	18.7	**0.5689**	**0.4884**	0.5150	0.4674	0.1827	**0.7216**

Qualitative Results. Such superiority is further corroborated by Figure 3. Our Narrative Adaptive Memory successfully links the prompt description with the exact person in previous frames, maintaining subject consistency. While our baseline LongLive (Yang et al., 2025a), and Self Forcing (Huang et al., 2025b) fail to connect visual cues with semantic instructions, resulting in disharmony between prompt-described transitions and actual video progression, and inability to maintain long-term consistency in cases where subjects disappear and reappear (as shown in Figure 1). Other approaches exhibit more severe error accumulation, with subject inconsistency in SkyReels-V2 (Chen et al., 2025b) and color drifting in FramePack. A user study with 20 participants (Table 3) further supports the effectiveness of our approach in visual quality, instruction following, and global consistency.

In terms of speed, MEMFLOW is 38× faster than SkyReels-v2 (Chen et al., 2025b), slightly faster than Self Forcing (Huang et al., 2025b), while slightly slower than LongLive (Yang et al., 2025a) due to memory operations, achieving significant memory performance with negligible speed sacrifice.

4.3 COMPARISONS FOR SINGLE-PROMPT GENERATION

Although not specifically trained for single-prompt generation, MEMFLOW shows superior performance on VBench (Huang et al., 2024b) for both 5s and 30s durations.

Short Video Generation. As shown in Table 2, MEMFLOW achieves the highest total score among open-source models of similar scale, with a clear advantage in semantic score attributed to prompt-aligned retrieval from Narrative Adaptive Memory. Despite an 8.6% speed reduction from memory operations compared with LongLive, our model still outperforms other methods with 18.7 FPS for real-time inference without degrading short-clip quality.

Long video generation. The superiority becomes more pronounced for 30-second videos (Table 4), where MEMFLOW shows consistent improvements across all metrics. Our adaptive memory provides more semantic-consistent context compared with local-window-only methods (Chen et al., 2025b; Huang et al., 2025b), context compression (Zhang & Agrawala, 2025), or first-chunk anchoring (Yang et al., 2025a). As a result, MEMFLOW maintains an advantage in long video generation quality with comparable performance on efficiency.

4.4 ABLATION STUDIES

We ablate our core components in a 60-second narrative task with five prompt switches (Table 5).

Narrative Adaptive Memory. We ablate different memory mechanism by comparing (i) *Frame Sink* (row 1): retaining KV cache from the first chunk (Yang et al., 2025a); (ii) *w/o Semantic Retrieval* (row 2): maintaining the entire memory bank without retrieval; and (iii) MEMFLOW (row 5): our full model. Results show that the *Frame Sink* baseline, by exclusively anchoring to the first chunk, fails to maintain consistency for subjects or scenarios that emerge in later segments and struggles to align with transitions depicted in the script. Conversely, maintaining the entire global bank (row 2) introduces excessive identity redundancy. The presence of redundant information distracts the attention mechanism, preventing the model from isolating and leveraging the most pertinent historical cues for consistent synthesis, leading to a significant drop in cross-segment consistency. Our retrieval-based NAM establishes precise dependencies across context, ensuring stable narrative transitions and identity persistence even during frequent subject switching.

Sparse Memory Activation. To evaluate SMA, we compare our full model against a variant where SMA is disabled (row 3). While both configurations produce consistent videos, SMA offers an optimal trade-off between generative quality and efficiency. By selectively activating the memory, SMA boosts the inference throughput from 17.6 to 18.7 frames per second (FPS). Furthermore, by implicitly filtering out less relevant history, MEMFLOW achieves more precise narrative transitions.

Memory-Forced Streaming Tuning. Removing MST (row 4) results in a significant deterioration across all visual and semantic consistency metrics. This drop suggests that exposing the model to multiple prompt switches—particularly challenging cases involving subject re-appearance—is essential. This strategy not only teaches the model to leverage retrieved history effectively but also enhances its capacity to manage complex content transitions. Consequently, MEMFLOW achieves superior performance across all memory-centric benchmarks.

As shown in Figure 4, *Frame Sink* preserves continuity only for initial subjects but collapses on later ones. In contrast, our model captures the relations between existing and emerging subjects, achieving superior semantic alignment on switched prompts, especially beyond 30 seconds.

4.5 ANALYSIS OF MEMORY RETRIEVAL ACCURACY

We verify retrieval effectiveness using narrative scripts where the primary subject is introduced (Segments 2-3), disappears (Segment 4), and reappears (Segments 5-6). We measure the retrieval hit rate—the probability that at least one retrieved KV pair originates from the injection phase when generating reappearance segments. Our results show the average retrieval accuracy reaches **75.58%**, confirming effective retrieval of semantically aligned context even after subject disappearance.

5 CONCLUSION

In this work, we introduce MEMFLOW, a memory mechanism for equipping interactive long video generation with long-range consistency without severe efficiency degradation. To maintain long-term memory for narrative coherence, we design Narrative Adaptive Memory for dynamic retrieval of semantic-aligned context via textual query. We also introduce Sparse Memory Activation for balancing the memory-efficiency trade-off by relevance-gated memory filtering. Our model achieves 18.7 FPS inference on a single NVIDIA H100 GPU, and supports interactive video generation while maintaining consistency, visual quality and narrative coherence even under complex narrative transitions and character switching.

AI USE STATEMENT

Gemini 2.5 Pro was used to construct the tailored prompt dataset for Memory-Forced Streaming Tuning, as described in Section 4.1. Large language models were also used for grammar checking and text polishing.

REFERENCES

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2025a.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model. *CoRR*, abs/2504.13074, 2025b.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *arXiv preprint arXiv:2505.21996*, 2025c.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *CVPR*, pp. 17702–17711, 2025.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *ICLR*, 2025.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *CoRR*, abs/2503.19325, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *CoRR*, abs/2501.00103, 2025.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, et al. Slowfast-vgen: Slow-fast learning for action-driven long video generation. *arXiv preprint arXiv:2410.23277*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. Filmaster: Bridging cinematic principles and generative ai for automated film generation. *arXiv preprint arXiv:2506.18899*, 2025a.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024a.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024b.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. In *arXiv*, 2024c.

Jiaxiu Jiang, Wenbo Li, Jingjing Ren, Yuping Qiu, Yong Guo, Xiaogang Xu, Han Wu, and Wangmeng Zuo. Lovic: Efficient long video generation with context compression. *arXiv preprint arXiv:2507.12952*, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025.

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

Kuaishou. Kling video model. `https://kling.kuaishou.com/en`, 2024.

Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*, pp. 468–485. Springer, 2024.

OpenAI. Sora: Creating video from text, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Wenqi Ouyang, Zeqi Xiao, Danni Yang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Tokensgen: Harnessing condensed tokens for long video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18197–18206, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4172–4182, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pp. 8748–8763, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *ArXiv*, abs/2512.14614, 2025. URL `https://api.semanticscholar.org/CorpusID:283909443`.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. MAGI-1: autoregressive video generation at scale. *CoRR*, abs/2505.13211, 2025.

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.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5417–5432, 2025.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory, 2026. URL https://arxiv.org/abs/2604.08995.

Xunzhi Xiang, Yabo Chen, Guiyu Zhang, Zhongyu Wang, Zhe Gao, Quanming Xiang, Gonghu Shang, Junqi Liu, Haibin Huang, Yang Gao, et al. Macro-from-micro planning for high-quality and parallelized autoregressive long video generation. *arXiv preprint arXiv:2508.03334*, 2025.

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. *arXiv preprint arXiv:2507.18634*, 2025a.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025b.

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

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b.

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

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

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025.

Shangjin Zhai, Zhichao Ye, Jialin Liu, Weijian Xie, Jiaqi Hu, Zhen Peng, Hua Xue, Danpeng Chen, Xiaomeng Wang, Lei Yang, et al. Stargen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26822–26833, 2025.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *CoRR*, abs/2504.12626, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Pretraining frame preservation in autoregressive video memory compression. *arXiv preprint arXiv:2512.23851*, 2025a.

Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T Freeman, and Hao Tan. Test-time training done right. In *arXiv*, 2025b.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequences. In *ICLR*, 2025.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8806–8817, 2024.

A Overall Structure

In this appendix, we first provide preliminaries to supplement the foundation of our approach in Sec B, including the autoregressive video diffusion framework and the distribution datching distillation (DMD) training method. Next, more details of training configurations are provided in Sec C. We then describe the evaluation metrics for multi-prompt generation in Sec D. Then in Sec F, we give a more detailed discussion of ablation study for memory mechanisms, activated memory capacity and configuration of Narrative Adaptive Memory (NAM) & Sparse Memory Activation (SMA). Afterward, we present more quantitative and qualitative comparisons to validate the effecitive of MemFlow in Sec E. Finally, in Section G, limitations are discussed. We also make **a demo video** and **a local website** to showcase the dynamic effects of our interactive long video generation in supplementary material.

B Preliminaries

In this section, we establish the foundational concepts upon which our work is built. We first review the architecture of autoregressive video diffusion models, followed by an overview of the Distribution Matching Distillation loss that facilitates video-level training.

B.1 Autoregressive Video Diffusion Models

The generation of video content can be approached through a hybrid framework that synergizes the sequential modeling of autoregressive methods with the high-fidelity synthesis of diffusion models, which is known as autoregressive video diffusion models. Given a video with N frames, $x^{1:N} = (x^1, \ldots, x^N)$, the joint probability distribution over this sequence is decomposed via the chain rule of probability into a product of conditional distributions:

$$p(x^{1:N}) = \prod_{i=1}^{N} p(x^i | x^{<i}). \tag{6}$$

Each conditional factor $p(x^i|x^{<i})$ is then instantiated by a distinct diffusion process. This process synthesizes a new frame x^i by iteratively refining a sample from Gaussian noise distribution, conditioned on the previously generated frames $x^{<i}$. Such a formulation elegantly captures the temporal dependencies inherent in video data while leveraging the power of diffusion models to produce high-quality visual content. Although contemporary methods (Yin et al., 2025; Teng et al., 2025) often generate multi-frame chunks to improve efficiency, for notational simplicity, we will hereafter refer to each generated unit, whether a single frame or a chunk, as a "frame".

B.2 Distribution Matching Distillation (DMD)

Autoregressive paradigm performs self-rollouts, where the model generates complete video samples directly from the inference-time model distribution. This enables the application of video-level objective functions designed to minimize the discrepancy between the distribution of generated videos and that of real-world videos.

Distribution Matching Distillation (DMD) (Yin et al., 2024a; 2025) is one such objective that effectively minimizes the reverse Kullback-Leibler (KL) divergence between the two distributions. The optimization is driven by the difference between their respective score functions. The gradient for the DMD loss can be formulated by approximating the gradient of the reverse KL divergence, averaged over time t:

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{DMD}} &\triangleq \mathbb{E}_t \left[\nabla_\theta \text{KL} \left(p_{\text{gen},t} \| p_{\text{data},t} \right) \right] \\ &\approx -\mathbb{E}_t \left[\int \left(s_{\text{data}}(y_t, t) - s_{\text{gen}}(y_t, t) \right) \nabla_\theta y_0 \, d\epsilon \right]. \end{aligned} \tag{7}$$

where $y_0 = G_\theta(\epsilon)$ is the output of the generator G_θ for a random Gaussian noise input ϵ, and $y_t = \Psi(y_0, t)$ represents the result of applying the forward diffusion process Ψ to y_0. The terms s_{data} and s_{gen} denote the score functions of the true data distribution $p_{\text{data}}(x_t)$ and the generator's

Table A: **Qualitative analysis of different memory mechanisms** under double-prompt 10-second setting. "w/o Memory" means only attending to the local attention window, "Frame Sink" refers to keeping the kv cache from the first chunk as memory (Yang et al., 2025a), "NAM" adopts the whole memory bank without filtering, and "NAM+SMA" is our full model which further compresses memory by relevance-gated selection.

Memory Mechanism	Subject Consistency ↑	Background Consistency ↑	CLIP Score ↑	Throughput (FPS) ↑
w/o Memory	93.88	91.58	**26.11**	**23.5**
"Frame Sink"	**97.58**	**98.69**	25.73	<u>20.3</u>
NAM	<u>97.34</u>	<u>98.48</u>	<u>25.92</u>	17.6
NAM+SMA	97.01	98.46	25.84	18.7

output distribution $p_{\text{gen}}(x_t)$, respectively. A significant benefit of this approach is that, once the score functions are trained, the DMD optimization process for the generator G_θ becomes entirely self-contained, requiring no further access to the ground-truth data (Yin et al., 2024a).

C Training Configurations

C.1 Training Pipeline

Our training pipeline is engineered to optimize for sample efficiency, inference speed, and long-term fidelity. To achieve this, we integrate several key strategies, building upon the insights of prior work (Yin et al., 2025; Huang et al., 2025b; Yang et al., 2025a).

Sample-Efficient Training. To maximize learning from each data sample, our model adopts a training scheme where all input frames are supervised at every iteration (Huang et al., 2025b). This dense supervision ensures that the model rapidly learns the underlying data distribution. Concurrently, for efficient inference, we employ a key-value (KV) caching mechanism (Yin et al., 2025). This allows the model to reuse computations from previous steps, significantly accelerating the autoregressive generation process without redundant calculations.

Asymmetric Distillation. While KV caching enhances efficiency, we further accelerate the generation process through knowledge distillation. Specifically, we employ an asymmetric distillation strategy (Yin et al., 2025; Huang et al., 2025b), where our causal student model learns from a powerful, pre-trained teacher diffusion model equipped with bidirectional attention. The knowledge transfer is facilitated by Distribution Matching Distillation (DMD) (Yin et al., 2024a), a few-step distillation technique. This approach allows our student model to approximate the teacher's high-quality output distribution while retaining a much faster, causal architecture.

Streaming Long Tuning. A critical challenge in autoregressive generation is the accumulation of errors over long sequences, as minor imperfections in early frames can cascade into significant degradation later on (Zhang & Agrawala, 2025). To address this, we adopt the training strategy following LongLive (Yang et al., 2025a) that aligns the training and inference conditions. During training, the model is required to generate extended sequences by conditioning on its own previously generated—and therefore imperfect—outputs. Supervision is applied across this entire self-generated rollout. This process explicitly exposes the model to its own potential failure modes and progressively degraded contexts, enabling it to be more robust. This technique effectively mitigates error accumulation, leading to substantially improved fidelity and temporal consistency in long video generation.

C.2 Implementation Details

MemFlow is built on Wan2.1-T2V-1.3B (Wan et al., 2025), which produces 5s clips at 16 FPS and 832×480 resolution. Our training pipeline can be divided into two stages: We first adapt the pretrained model into a few-step causal-attention model equipped with our specially designed memory in a Self Forcing DMD manner (Huang et al., 2025b); then perform streaming long tuning on 60s sequences for improving long-term fidelity and consistency (Yang et al., 2025a).

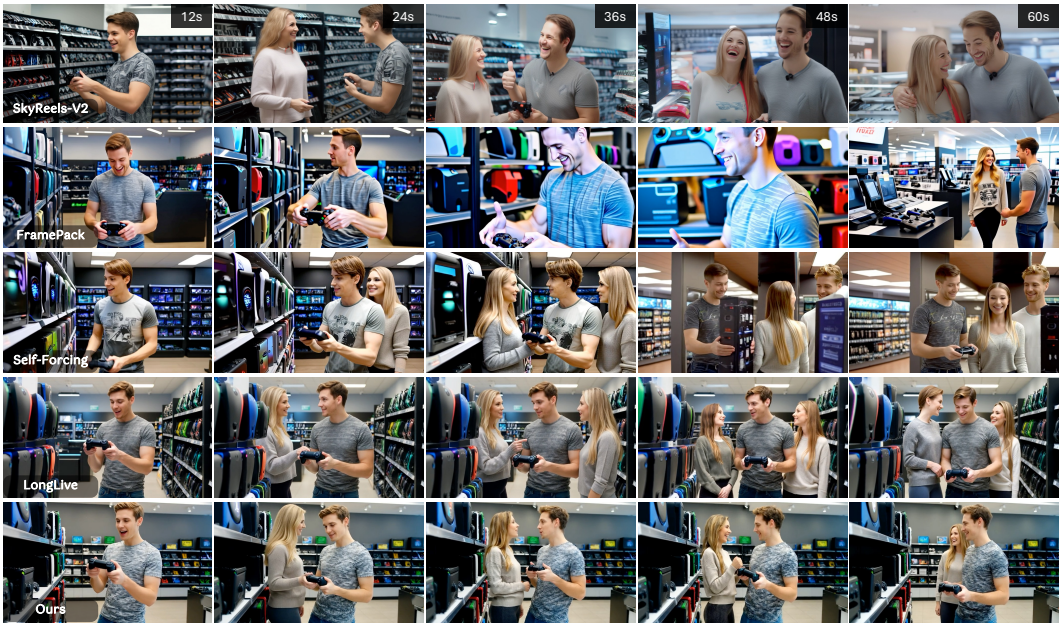

Figure A: **Qualitative comparisons under multi-prompt 60-second setting** with representative long video generation models, where MEMFLOW outperforms other alternatives in narrative coherence and subject consistency, without drifting or duplicated characters.

DMD. In the first stage, we adapt the pretrained Wan2.1-T2V-1.3B into a chunk-wise autoregressive (AR) causal attention model on a filtered and LLM-extended version of VidProM data (Wang & Yang, 2024). First, we conduct an ODE initialization as Self Forcing does. Then we train the model with DMD, while enabling our Narrative Adaptive Memory by equipping the original local context window (9 latent frames) with a memory bank of 3 latent frames. The final capacity of our memory is also set to 3 latent frames after sparse activation. The few-step diffusion process employs a uniform 4-step schedule $[t_4, t_3, t_2, t_1] = [1000, 750, 500, 250]$.

Streaming Tuning. The second stage uses a switch-prompt dataset from LongLive (Yang et al., 2025a), containing pairs of prompts with each containing two prompts, and the second adding exactly one new action/object for the existing main subject in the first prompt. Then we can perform streaming long tuning on every 60s sequence that includes a single prompt switch, with the switch time sampled uniformly from 5s to 55s. During training, each iteration continues the model's own rollout by generating the next 5s video clip until a maximum length of 60s is reached, supervised by Wan2.1-T2V-14B as the teacher. Specifically, in each iteration, the student model samples a short video clip (e.g., 5s), and the teacher provides supervision exclusively on this newly generated clip via DMD loss, while the previously generated clips remain frozen and serve as conditioning context. Our memory bank is updated before each iteration of autoregressive rollout, allowing the model to retrieve relevant historical information before generating the next clip. The training is conducted on 32 GPUs with two samples per GPU (global batch size is 64), and the first DMD stage lasts for 700 steps while the second streaming tuning stage lasts for 3000 steps. We use AdamW optimizer with learning rates of $lr = 1.0 \times 10^{-5}$ for actor and $lr_{critic} = 2.0 \times 10^{-6}$ for critic. LoRA (Hu et al., 2022) with $rank = 256$ is adopted throughout the streaming long tuning procedure.

D EVALUATION METRICS FOR MULTI-PROMPT GENERATION

We describe the memory-centric evaluation metrics used in our multi-prompt experiments. Given a 60-second video composed of six 10-second segments (each corresponding to one prompt), we compute the following:

Visual Drift. Following FramePack (Zhang & Agrawala, 2025), we measure quality degradation over time by computing the aesthetic score using the LAION predictor (Schuhmann et al., 2022) for



Figure B: **Qualitative comparisons under multi-prompt 60-second setting** with representative long video generation models, where MEMFLOW outperforms other alternatives in narrative coherence and subject consistency, without drifting or duplicated characters.

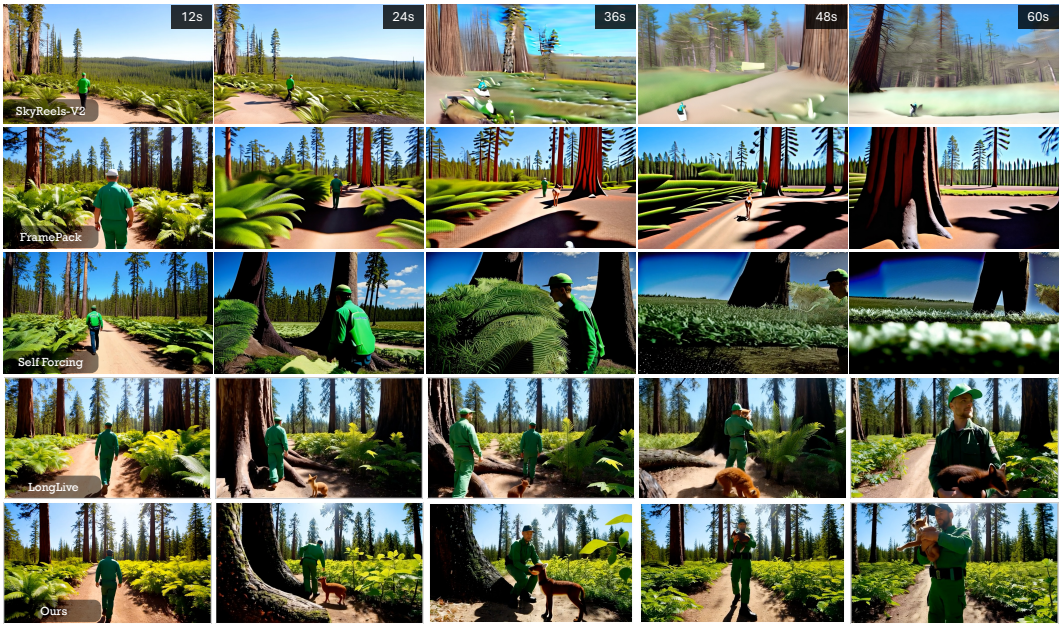

Figure C: **Qualitative comparisons under multi-prompt 60-second setting** with representative long video generation models, where MEMFLOW outperforms other alternatives in narrative coherence and subject consistency, without drifting or duplicated characters.

each segment and reporting the absolute difference between the first and last segments. Lower drift indicates better resistance to error accumulation during long autoregressive rollouts.

Cross-segment Consistency. We evaluate the visual consistency of subjects and backgrounds across prompt-segmented video segments. Specifically, we use SAM-3 (Carion et al., 2025) to segment

Table B: **VBench scores across quality dimensions** under single-prompt 5-second and 30-second settings of MEMFLOW and LongLive (Yang et al., 2025a).

Model	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Quality Score
MEMFLOW (5 s)	97.71	97.63	99.13	98.89	58.33	68.57	72.12	85.95
LongLive (5 s)	97.44	97.20	99.28	98.90	41.67	66.55	70.00	83.96
MEMFLOW (30 s)	99.07	97.52	99.13	99.26	59.26	64.58	72.73	85.92
LongLive (30 s)	98.98	97.57	99.35	99.29	34.35	64.31	72.21	83.98

Table C: **VBench scores across semantic dimensions** under single-prompt 5-second and 30-second settings of MEMFLOW and LongLive (Yang et al., 2025a).

Model	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency	Semantic Score
MEMFLOW (5 s)	98.97	97.87	80.00	97.49	92.94	70.42	21.53	25.39	27.69	81.90
LongLive (5 s)	95.56	87.58	82.00	94.30	82.88	55.60	20.58	24.09	26.65	80.15
MEMFLOW (30 s)	92.87	94.59	78.60	93.70	80.87	72.69	20.91	24.24	26.56	78.87
LongLive (30 s)	93.99	71.94	79.00	88.89	82.13	60.69	19.37	21.19	21.19	74.80

foreground subjects and background regions from the median frame of each segment. For each segmented entity, we extract CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) features. For segments that share the same identity (as specified by the narrative script), we cluster their features and compute pairwise cosine similarity. For transition segments where the script specifies that a subject should have disappeared, we apply a penalty if the original entity erroneously persists in the generated video. The final score is the average across all segment pairs, separately reported for subjects and scenes.

Semantic Consistency. This metric evaluates how well the generated video aligns with the narrative script at the semantic level. It consists of two sub-metrics:

- *Prompt Alignment*: For each segment, we compute the mean ViCLIP (Wang et al., 2023) cosine similarity between the generated video segment and its corresponding text prompt, measuring per-segment instruction following.
- *Transition Alignment*: At each prompt boundary, we verify whether the subject presence/absence matches the script specification. For example, if the script dictates that a character disappears in segment 4 and reappears in segment 5, we check whether the generated video correctly reflects this transition. This metric specifically evaluates the model's ability to handle complex narrative plots involving subject injection and removal.

E MORE COMPARISONS.

E.1 COMPARISONS FOR MULTI-PROMPT GENERATION

Since our memory mechanism is designed for interactive long-form videos with multiple prompts, we compare MEMFLOW's abilities with representative long video generation models. For fair comparison, we adapt existing methods—including SkyReels-V2 (Chen et al., 2025b), FramePack (Zhang & Agrawala, 2025), and Self Forcing (Huang et al., 2025b)—for multi-prompt video generation by switching prompts during the autoregressive synthesis. Note that LongLive (Yang et al., 2025a) inherently supports generating videos with interactive instructions. Following LongLive (Yang et al., 2025a), we customize 100 groups of narrative scripts, with each consisting of 6 successive 10-second prompts for a total of 100 videos lasting for 60 seconds.

We use metrics from VBench-Long (Huang et al., 2024c) for assessing visual quality of all generated videos, among which the two dimensions of consistency and aesthetic are highlighted for comparison of long-range consistency in subjects, background and visual aesthetics. Table F shows that MEMFLOW achieves the best quality score among all methods, verifying its comprehensive competitiveness in perceptual quality. In terms of consistency score, our method outperforms all other models except Framepack (Zhang & Agrawala, 2025), with Framepack tending to synthesize videos



Figure D: **Qualitative analysis of different memory mechanisms** under multi-prompt 60-second setting. "w/o Memory" means only attending to the local attention window, "Frame Sink" refers to keeping KV cache from the first chunk as memory, "NAM" adopts the whole memory bank without filtering, and "NAM+SMA" is our full model which compresses memory by relevance-gated selection.



Figure E: **Qualitative analysis of different memory mechanisms** under multi-prompt 60-second setting. "w/o Memory" means only attending to the local attention window, "Frame Sink" refers to keeping KV cache from the first chunk as memory, "NAM" adopts the whole memory bank without filtering, and "NAM+SMA" is our full model which compresses memory by relevance-gated selection.

with reduced inter-frame dynamics, which definitely shows superiority in "consistency". Thus the result can still demonstrate our superior performance in global consistency with the specially designed memory. Our advantage in aesthetic score also highlights the effectiveness of our method in mitigating error accumulation during long rollouts.

For text alignment with interactive prompts, videos are segmented according to prompt boundaries for evaluating clip-wise semantic adherence. The CLIP score (Radford et al., 2021) between each video clip and its corresponding text is calculated at 10-second intervals. The results demonstrate outstanding prompt adherence and narrative coherence, particularly when the video is extrapolated to longer durations, owing to the model's ability to establish long-term contextual associations.



Figure F: **Qualitative analysis of different memory mechanisms** under multi-prompt 60-second setting. "w/o Memory" means only attending to the local attention window, "Frame Sink" refers to keeping KV cache from the first chunk as memory, "NAM" adopts the whole memory bank without filtering, and "NAM+SMA" is our full model which compresses memory by relevance-gated selection.

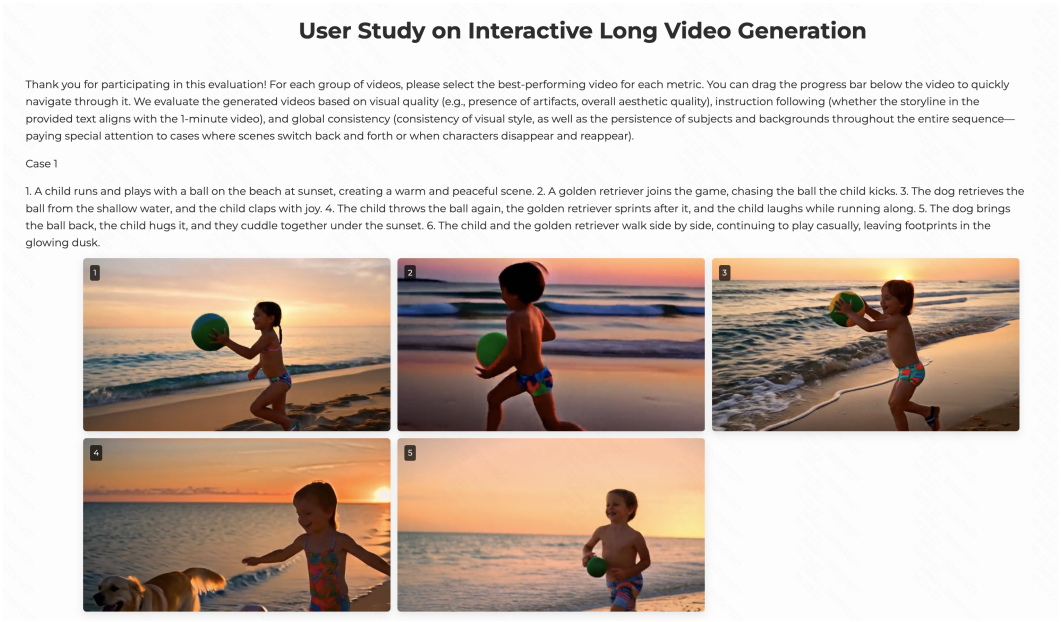

Figure G: User study instruction screenshots.

Such superiority is further corroborated by the qualitative results in Figure A. Our Narrative Adaptive Memory successfully links description in the prompt, "a woman in a casual sweater", with the exact person in previous frames, thus maintaining the subject consistency. While our baseline, LongLive (Yang et al., 2025a), fails to make connections between visual cues and semantic instructions, therefore continuously introduces new characters after prompt switching, exhibiting less temporal coherence and prompt compliance. Other approaches exhibit more severe error accumulation, with subject inconsistency in SkyReels-V2 (Chen et al., 2025b) and color drifting in FramePack. Self Forcing (Huang et al., 2025b) also suffers from similar problems with LongLive, showing misalignment between prompt scripts and narrative progression across clips as characters are repeatedly introduced into the ongoing scene.

Table D: **Quantitative analysis of different memory mechanisms** under multi-prompt 60-second setting. "w/o Memory" means only attending to the local attention window, "Frame Sink" refers to keeping KV cache from the first chunk as memory (Yang et al., 2025a), "NAM" adopts the whole memory bank without filtering, and "NAM+SMA" is our full model which compresses memory by relevance-gated selection.

Memory Mechanism	Subject Consistency ↑	Background Consistency ↑	Throughput (FPS) ↑	CLIP Score ↑					
				0–10 s	10–20 s	20–30 s	30–40 s	40–50 s	50–60 s
w/o Memory	94.41	95.15	**23.5**	26.74	25.10	24.60	23.61	24.23	24.14
"Frame Sink" (Yang et al., 2025a)	97.66	96.20	20.3	**26.63**	**25.77**	**24.65**	23.99	24.52	24.11
NAM+SMA	98.01	**96.70**	18.7	26.31	24.70	23.94	24.13	24.90	24.22
NAM	**98.05**	96.57	17.6	26.50	25.30	24.42	**24.23**	**24.96**	**24.28**

Table E: **Quantitative comparison under multi-prompt 60-second setting** with representative long video generation models, where SkyReels-V2 (Chen et al., 2025b), Self-Forcing (Huang et al., 2025b) and FramePack (Zhang & Agrawala, 2025) are adapted for the task by directly switching prompts.

Model	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Quality Score
SkyReels-V2 (Chen et al., 2025b)	95.65	95.96	99.23	56.83	63.25	81.55
Self-Forcing (Huang et al., 2025b)	97.18	96.15	98.98	58.46	72.43	83.94
LongLive (Yang et al., 2025a)	98.11	96.72	99.12	59.90	70.34	84.28
FramePack (Zhang & Agrawala, 2025)	98.18	96.81	99.51	59.44	69.86	84.40
MEMFLOW	98.01	96.70	99.04	63.75	70.57	**85.02**

E.2 VBENCH SCORES ACROSS ALL DIMENSIONS

As in Tables 1, 2, and 4, we evaluate our method for multi-prompt and single-prompt video generation on VBench (Huang et al., 2024b) (or VBench-Long (Huang et al., 2024c)), a comprehensive benchmark that contains 16 metrics covering various aspects. All values are computed with the official standardized evaluation scripts.

For single-prompt generation, we use all 946 official prompts and all 16 metrics as reported in Tables 2 and 4, covering durations of 5s and 30s. We assess the performance of 5-second videos on Vbench (Huang et al., 2024b) while for 30-second videos, we evaluate on VBench-Long (Huang et al., 2024c). The detailed scores across all dimensions are reported in Table B and Table C. Compared with the state-of-the-art model LongLive (Yang et al., 2025a), ours excels in terms of semantic alignment, evidenced by the especially high scores in multiple objects, spatial relationship, and scene. For visual quality, MEMFLOW is comparable to LongLive (Yang et al., 2025a) as in Table B, with notably higher dynamic degree score, indicating more dynamic motion in the generated videos.

For multi-prompt generation, we follow LongLive (Yang et al., 2025a) to customize 100 groups of narrative scripts, with each consisting of 6 successive 10-second prompts for a total of 100 videos lasting for 60 seconds. The metrics from VBench-Long (Huang et al., 2024c) including subject consistency, background consistency, motion smoothness, aesthetic quality, and imaging quality, are employed for assessing visual quality shown in Table E. MEMFLOW shows an overall advantage demonstrated by the total quality score, and especially outstanding performance in aesthetic quality due to our anti-drifting effect of long-range memory.

E.3 USER STUDY

We conduct a user study with 20 participants to compare our method against the representative long video generation models (SkyReels-V2 (Chen et al., 2025b), FramePack (Zhang & Agrawala, 2025), Self Forcing(Huang et al., 2025b), and LongLive (Yang et al., 2025a)). It includes human preference rates in visual quality, instruction following, and global consistency. For each group of videos, participants were shown all five videos together with the corresponding prompt and asked to select the model with the best performance on the three dimensions. Detailed instructions are shown in Fig G. The results further support the effectiveness of our approach.

Table F: **Quantitative comparison under multi-prompt 60-second setting** with representitive long video generation models, where SkyReels-V2 (Chen et al., 2025b), Self Forcing (Huang et al., 2025b) and FramePack (Zhang & Agrawala, 2025) are adapted for the task by directly switching prompts. All scores are measured over the whole sequence, except for the CLIP score, which is computed at intervals aligned with the prompt switching.

Method	Quality Score ↑	Consistency Score ↑	Aesthetic Score ↑	CLIP Score ↑					
				0–10 s	10–20 s	20–30 s	30–40 s	40–50 s	50–60 s
SkyReels-V2 (Chen et al., 2025b)	81.55	94.72	56.83	25.31	23.40	22.50	21.62	21.67	20.91
Self Forcing (Huang et al., 2025b)	83.94	95.74	58.45	26.24	24.87	23.46	21.92	22.05	21.07
LongLive (Yang et al., 2025a)	84.28	96.05	59.89	**26.63**	**25.77**	24.65	23.99	24.52	24.11
FramePack (Zhang & Agrawala, 2025)	84.40	**96.77**	59.44	26.51	22.60	22.18	21.53	21.98	21.62
MEMFLOW	**85.02**	96.60	**61.07**	26.49	**25.77**	**24.89**	**24.99**	**25.49**	**24.64**

F MORE ABLATIONS

F.1 MEMORY MECHANISMS

Interactive Multi-Prompt Video Generation. We perform ablation studies on different memory mechanisms based on our framework, on a 60-second interactive multi-prompt video generation setting with five prompt switches. We ablate by comparing (i) w/o memory: conditioned only on local context window; (ii) "Frame Sink": additionally retain the KV cache from the first chunk; (iii) Narrative Adaptive Memory (NAM): maintain our dynamic memory bank; and (iiii) Narrative Adaptive Memory and Sparse Memory Activation (NAM+SMA): our full model. "Frame Sink" is utilized by LongLive (Yang et al., 2025a), allowing direct comparison with its model. The configuration of w/o memory is also implemented on the baseline of LongLive by removing the sink latent frames. Table D highlights the effectiveness of our NAM, which consistently outperforms others in maintaining temporal consistency and semantic coherence. The retrieval-based memory establishes intrinsic dependencies across contexts, enabling stable narrative transitions even under subject insertion or switching. With SMA, inference efficiency improves from 17.6 to 18.7 frames per second with minimal quality degradation. As shown in Figure F, removing memory causes abrupt scene transitions, while "Frame Sink" preserves continuity only for initial subjects but collapses on later ones. In contrast, our model captures the relations between existing and emerging subjects, achieving superior semantic alignment on switched prompts, especially as the video extends beyond 30 seconds.

Single-Prompt Video Generation. We also ablate different memory mechanisms on 10-second videos with a single prompt switch at the 5-second following LongLive (Yang et al., 2025a), where we preserve the main subject along the whole video and only add new action, but no new object for the existing scene. In Table A, the metrics of "Subject Consistency" and "Background Consistency" are from Vbench (Huang et al., 2024b), and CLIP Score is computed at intervals aligned with the prompt switching. Thus the final score is the average of the two pairs of textual and visual similarity between video clip and its corresponding text at 5-second intervals.

The results show that in the 10-second setting, "Frame Sink" achieves the best subject and background consistency among the variants and our "NAM" ranks the second place. This is expected, since the main subject remains unchanged throughout the entire 10-second duration in our evaluation, and retaining the KV cache from the first chunk provides a stable and clean historical signal. In contrast, our NAM may not appear to be as stable as this hard-coded strategy under such a simple scenario. We also note that the configuration of "w/o memory" outperforms other baselines on CLIP Score, which is acceptable that the generation is not interrupted by redundant context and is focused on narrative generation. Such phenomenon can also be observed in video identity customization, which can be viewed as a normal trade-off between consistency and prompt alignment.

Taken together, these observations indicate that while the "Frame Sink" mechanism holds an advantage when the main subject stays unchanged—benefiting from the clean and reliable KV cache obtained from the first chunk—it struggles once the narrative undergoes more complex transitions such as scene switching or character insertion. In these cases, our approach demonstrates clear superiority (Table A).

Our qualitative comparisons further validate this conclusion. As illustrated in Fig D and Fig E, without any memory mechanism, both foreground and background drift substantially after the camera

moves to a different region (e.g., toward the gifts under the christmas tree) and then returns to the original scene, failing to match the initial appearance. When solely relying on the first chunk as memory, the overall scene and the main subject (the man in the green sweater) introduced in the first chunk remain consistent across the entire video. However, the inserted subject introduced later in the video (the woman in the red sweater) undergoes severe identity changes, showing that the hard-coded "Frame Sink" strategy cannot adapt to scene transitions or the insertion of new objects or characters.

In contrast, our method maintains a retrieval-based memory bank, enabling different video chunks to reference distinct, relevant historical cues. This design allows the model to effectively handle scene changes, and multi-character dynamics, thereby achieving robust long-range contextual consistency.

G Limitations and Future Works

As an efficient fine-tuning scheme, the ultimate performance of MemFlow is inherently bounded by the capacity and quality of its pretrained base model. Our reliance on a self-supervised fine-tuning strategy, which obviates the need for additional curated real-video data, enhances both efficiency and scalability. However, this design choice also implies that the method has limited capacity to rectify systematic errors or biases inherited from the base model. Moreover, there is still an inability to maintain consistent quality throughout videos lasting half an hour or more, which needs future works to figure out how to overcoming this barrier. Additionally, as the generated video grows longer, the memory bank capacity increases accordingly. For very long sequences, offloading the memory bank to CPU may become necessary to avoid exceeding GPU memory limits, which could introduce additional latency during retrieval operations.