MEMID: MEMORIZING IDENTITIES IN STREAMING NARRATIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

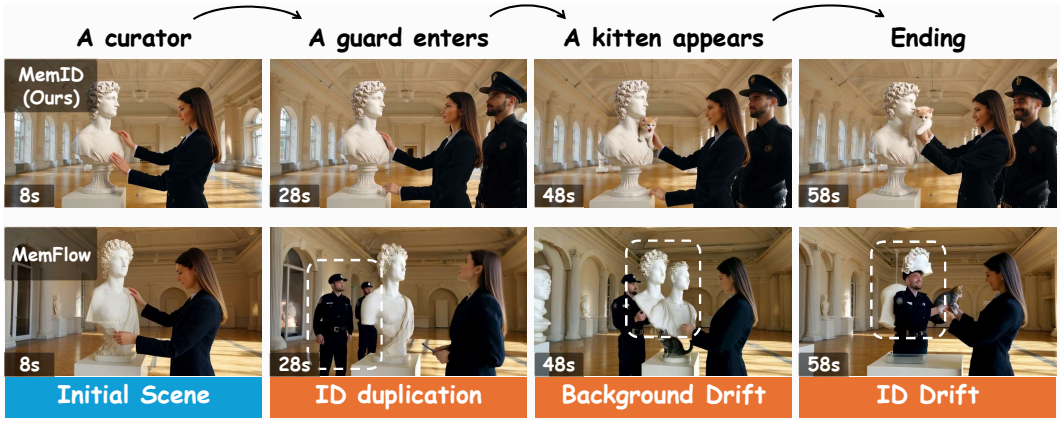


Figure 1: MemID is an identity-aware autoregressive video diffusion model that preserves identities across streaming long-video narratives, achieving consistent character generation compared with the baseline MemFlow (Ji et al., 2025).

ABSTRACT

Memory is central to streaming autoregressive (AR) video generation, where later events of a multi-prompt narrative must stay consistent with content that has fallen outside the bounded context window. Existing memory methods mainly focus on *what to remember*, yet they either overwrite early identity anchors through continuous memory updates or rely on pretrained LLMs/VLMs as external retrievers in inference. More importantly, they overlook *where to render* each identity: even with correct identity references, overlapping attention between different identities causes ID fusion, while fragmented attention within the same identity leads to ID duplication. We propose *MemID*, an identity-aware streaming AR video generation framework that addresses both aspects. For what to remember, MemID introduces an *ID-preserving Memory* that compresses each prompt-conditioned event into representative frames, freezes it after the event ends, and retrieves only prompt-relevant entries for current generation. For where to render, MemID is trained with *ID-aware Attention Losses* that separate different identities while encouraging compact and temporally coherent attention for each identity. Extensive experiments demonstrate that MemID substantially improves long-range identity consistency and maintains accurate instance counts in multi-character narratives, while preserving competitive generation quality. Code and models will be publicly released.

1 INTRODUCTION

Autoregressive (AR) video diffusion models (Huang et al., 2025b; Henschel et al., 2025; Yang et al., 2025; Yesiltepe et al., 2026) have made remarkable progress in streaming text-to-video generation, producing long videos chunk by chunk under a sequence of prompts. A long video is usually a story, in which characters enter, leave, and come back later. But AR models attend only to a bounded context window, so a character who has been away for too long is forgotten and comes back looking like

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

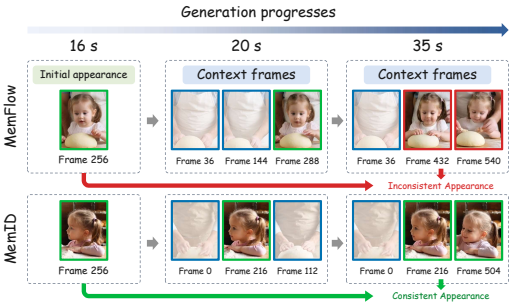


Figure 2: MemFlow overwrites old frames in memory bank, causing appearance drift, while MemID retains memory from the earlier frames for better consistency.

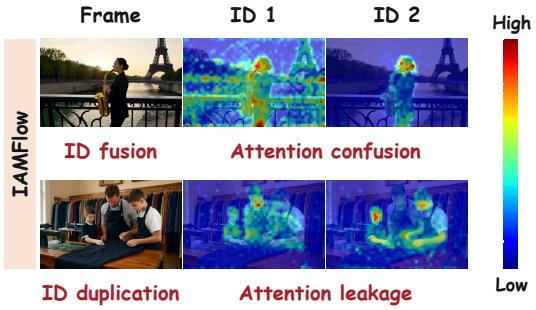


Figure 3: IAMFlow exhibits attention mismatch during generation, manifested as identity fusion and duplication, thereby resulting in severe identity inconsistency.

someone else, as shown in Fig. 1. Memory is therefore the central problem in streaming narrative generation, and character identity is what it should preserve first.

To equip AR models with memory, existing methods compress past frames into a compact context (Zhang et al., 2025a; Jia et al., 2025; Huang et al., 2025a; Meng et al., 2025; Zhang et al., 2025b), retrieve them through 3D geometry (Li et al., 2025a; Zhang et al., 2025c; Wang et al., 2025; Yu et al., 2025; HunyuanWorld, 2025), or keep a bank of visual features to read from during generation (Ji et al., 2025; Chen et al., 2026). None of them is organized around characters: compression averages away the very details that tell one character from another, geometry indexes the scene rather than the people in it, and a memory bank that is rewritten as generation proceeds can replace a reliable early appearance with a later, already-drifted one, as shown in Fig. 2. IAMFlow (Liu et al., 2026) is the only method that targets identity directly, calling external LLMs and VLMs at inference to retrieve character references; but its generator is never trained to use what is retrieved, so retrieval errors pass straight through and ID fusion and duplication remain, as shown in Fig. 3.

We revisit memory in AR video generation from the perspective of character identity. We argue that preserving identity requires deciding both *what to remember* over time and *where to render* each identity in space. Existing methods address the former, but at the level of whole scenes rather than individual identities. Besides, even with correct identity references in memory, the retrieved features are useless if the identity is rendered in the wrong place, and the model still suffers from *ID fusion* and *ID duplication* in multi-character scenes. We attribute them to two attention mismatch patterns, *inter-ID attention confusion* and *intra-ID attention leakage*, where attention is entangled across different IDs or scattered over disconnected regions of the same ID, respectively.

To this end, we propose *MemID*, an identity-aware streaming AR video generation framework that addresses both what to remember and where to render based on a Self-Forcing-style streaming long-video generation framework (Huang et al., 2025b). For *what to remember*, MemID introduces an *ID-preserving Memory* that factorizes memory construction into event-wise compression and then text-video retrieval. Within each event, frame-level keys are clustered with K-Means, and each centroid is replaced by its nearest real frame, whose original token-level keys and values form one memory entry. This yields a compact memory without averaging away identity details. Once an event ends, its memory is frozen, so that identity anchors are never overwritten by subsequent generations. At each generation step, only the top-k entries most relevant to the current prompt are retrieved, allowing MemID to access distant events without linearly increasing the attention context. For *where to render*, MemID is trained with *ID-aware Attention Losses* that encourage spatially disentangled and temporally coherent identity-specific attention patterns. These losses consist of an *Exclusive Loss* and a *Tracking Loss*. Specifically, the *Exclusive Loss* penalizes spatial overlap between the attention maps of different identities, encouraging each identity token to attend to a distinct region and thereby mitigating identity fusion. Meanwhile, the *Tracking Loss* promotes temporal consistency by suppressing disconnected or spurious activations outside a temporally anchored region for each identity. By encouraging identity-specific attention to remain concentrated around its previously established region across frames, the Tracking Loss reduces identity duplication and prevents the same identity from being rendered at multiple locations. Together, these two objectives provide a spatial-annotation-free mechanism for controlling *where* each identity should appear. Extensive experiments demonstrate that MemID substantially improves long-range identity consis-

tency over existing streaming and memory-based approaches, while maintaining competitive visual quality and real-time inference.

The main contributions of our work are summarized as follows:

- We revisit memory in streaming AR video generation from the perspective of character identity, and identify that identity inconsistency arises not only from *what to remember* over time but also from *where to render* each identity in space, where inter-ID attention confusion and intra-ID attention leakage lead to ID fusion and ID duplication.
- We propose an *ID-preserving Memory* that combines event-wise memory compression with text-video context retrieval, which freezes representative identity anchors of completed events and exposes only a compact, prompt-relevant subset of the history to each step.
- We introduce *ID-aware Attention Losses*, consisting of an Exclusive Loss and a Tracking Loss, which teach the generator where to render each identity by assigning it a distinct and spatially coherent attention region.

2 Related Work

Streaming Multi-prompt Video Generation. Multi-prompt long video generation aims to produce videos that follow a sequence of textual prompts with smooth transitions and consistent content. Early approaches are largely training-free: built on pre-trained bidirectional diffusion models (Ho et al., 2020; Rombach et al., 2022), they generate short clips and stitch them together by manipulating latents, noise schedules, or attention across segments (Wang et al., 2023; Qiu et al., 2024; Kim et al., 2024; Cai et al., 2025; Lu et al., 2024; Kim et al., 2025). Since each segment comes from a fixed-length model, they do not scale to open-ended, interactive generation. Streaming methods instead generate videos autoregressively, extending the video chunk by chunk as new prompts arrive. Early AR extension conditions each clip on the last frames of the previous one (Blattmann et al., 2023; Henschel et al., 2025), but suffers from error accumulation and content drift over time. Self-Forcing (Huang et al., 2025b) distills a causal few-step generator from a bidirectional teacher and trains it on its own rollouts, enabling real-time streaming generation and inspiring a series of follow-up works (Liu et al., 2025; Shin et al., 2025; Li et al., 2025b; Yesiltepe et al., 2026). Building on this paradigm, LongLive (Yang et al., 2025) supports interactive multi-prompt generation with frame sinks and KV-recache for smooth prompt switching. However, these streaming models only attend to a bounded window of recent history, so content that falls outside this window, such as a character who has left the scene, is gradually forgotten.

Memory Mechanisms in Long Video Generation. Since full self-attention scales quadratically with sequence length, a long video model cannot keep its entire history in context, and must decide what to carry forward. Existing solutions follow three lines. *Compression-based* methods shorten the history through temporal downsampling, learned compressors, or inference-time fine-tuning (Zhang et al., 2025a; Jia et al., 2025; Huang et al., 2025a; Meng et al., 2025; Zhang et al., 2025b; Hong et al., 2024), which slows memory growth but averages away exactly the fine-grained appearance cues that tell one character from another. *Geometry-dependent* methods instead retrieve history through 3D structure or field-of-view overlap (Zhang et al., 2025c; Wang et al., 2025; Yu et al., 2025; HunyuanWorld, 2025; Li et al., 2025a); they suit static scene exploration, but rely on accurate camera poses and index the world by where the camera looked rather than by who was in the shot. *Memory-bank-based* methods store features of previously generated frames and retrieve them as references. Context-Forcing (Chen et al., 2026) and MemFlow (Ji et al., 2025) keep a capacity-constrained bank that is rewritten as generation proceeds, so a reliable early appearance can be displaced by a later, already-drifted one. IAMFlow (Liu et al., 2026) is closest to our setting, making retrieval identity-aware with pretrained LLMs and VLMs at inference; its generator, however, is not explicitly trained to utilize retrieved references, so parsing errors pass straight through, and even a correct reference does not say where to put the character. In contrast, MemID freezes the memory of each completed event so that subsequent generation cannot overwrite it, preserving identity anchors throughout the narrative.

Attention Control in Diffusion Models. Cross-attention maps encode a coarse spatial layout for each text token, which makes them a natural handle for deciding where a subject is rendered. Training-free methods exploit this by editing or steering these maps at inference time toward a

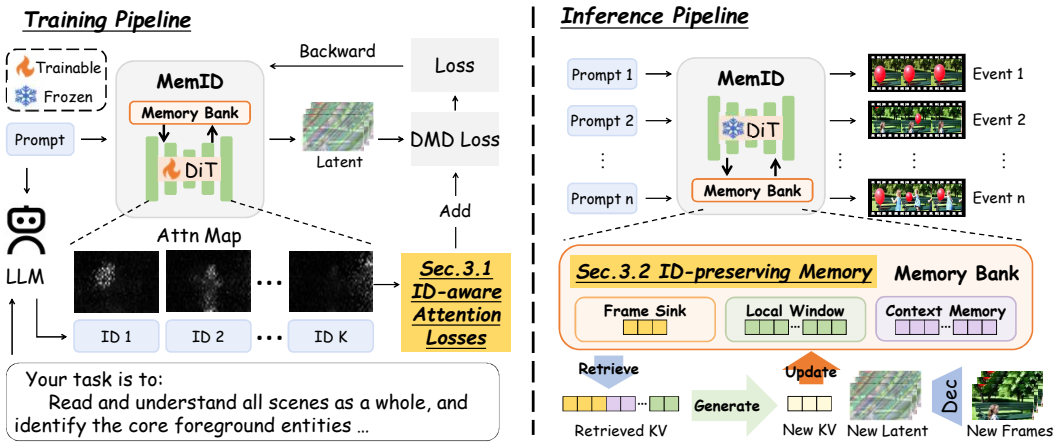


Figure 4: **Overview of MemID for multi-prompt streaming AR video generation.** Each prompt defines a narrative event. MemID consolidates completed events into persistent historical memory and regularizes identity-specific attention during training.

user-provided box or mask (Hertz et al., 2022; Chefer et al., 2023; Xie et al., 2023; Kim et al., 2023; Chen et al., 2024), an idea that extends to video through box trajectories (Jain et al., 2024; Yang et al., 2024; Ma et al., 2024), while GLIGEN (Li et al., 2023) instead learns to ground boxes with additional trainable layers. Closest to us is FastComposer (Xiao et al., 2025), which supervises cross-attention with segmentation masks so that concurrent subjects do not blend into each other. However, existing spatial-control methods typically rely on externally specified boxes, masks, or trajectories, either as training supervision or inference-time guidance. Neither paradigm integrates naturally with our streaming generation setting for identity preservation.

3 METHOD

The proposed *MemID* targets streaming, multi-prompt long-video generation with consistent character identities, where each character is expected to maintain a coherent appearance throughout the entire video. As illustrated in Figure 4, based on the Self-Forcing-style framework (Huang et al., 2025b; Yang et al., 2025), MemID is trained using *ID-aware Attention Losses* for spatially disentangling different character identities (Sec. 3.1) and integrates an event-wise *ID-preserving Memory* in both training and inference for maintaining character identities over long temporal horizons (Sec. 3.2). Below, we give the details of each part.

3.1 ID-AWARE ATTENTION LOSSES

During long-form video generation, unconstrained identity-specific attention can lead to severe ID inconsistency. We observe two characteristic failure modes. First, spatial overlap between the attention maps of different identities can entangle their visual attributes, resulting in *ID fusion*. Second, a single identity's attention may fragment into multiple disconnected regions, producing multiple spatial instantiations of the same identity and resulting in *ID duplication*. Thus, we propose *ID-aware Attention Losses*, consisting of *Exclusive Loss* to separate the attention maps of different identities and *Tracking Loss* to suppress fragmented regions outside the main region of each identity. We incorporate our ID-aware Attention Losses as a regularization term during training to guide the model toward more identity-consistent and spatially coherent attention patterns.

To compute the ID-aware Attention Losses, we first employ an off-the-shelf LLM to identify the entity IDs in the training prompts and extract their corresponding cross-attention maps from the diffusion model. The LLM is used only during training; no LLM or VLM is required at inference. Given a training prompt, it returns G entity spans, one per character. For the g-th span, we average the cross-attention weights over its text tokens and over attention heads, and rescale the resulting map to $[0, 1]$ by min-max normalization over the $H \times W$ spatial locations, giving one map $p_g^t \in \mathbb{R}^{H \times W}$ per identity and frame t. Min-max normalization places every identity on the same scale, so that both

losses respond to *where* an identity attends rather than to how much raw attention mass it happens to carry. Further details are provided in the supplementary material.

Exclusive Loss. ID fusion originates in a simple geometric fact: when two identities place attention mass on the same locations, their visual attributes are rendered into one shared region. The Exclusive Loss eliminates this shared support by penalizing pairwise attention overlap. For each unordered pair $i < j$, the pixel-wise product of their attention maps, $p_i^t(h,w)\, p_j^t(h,w)$, acts as a local overlap measure, which we average over all HW spatial locations, all $G(G-1)/2$ pairs, and all T frames of the current chunk:

$$\mathcal{L}_{\text{excl}} = \frac{2}{THWG(G-1)} \sum_t \sum_{i<j} \sum_{h,w} p_i^t(h,w) p_j^t(h,w). \tag{1}$$

This product is large only where both identities are simultaneously active and vanishes as soon as one of them is not, so the penalty is confined to genuinely contested locations, while the factor $\frac{2}{THWG(G-1)}$ keeps its scale independent of resolution and of the number of identities. The resulting gradient pushes identities apart in space and never acts on the shape of an individual attention map.

Tracking Loss. ID duplication, in contrast, is an intra-identity failure: the attention of one identity splits into several disconnected regions, each of which can be realized as a separate instance of that character. Separating identities does not help here; what is required is a commitment to one region per identity, held consistent over time. The Tracking Loss enforces this by designating a dominant region and suppressing whatever competes with it. Concretely, we build a coarse envelope by applying average pooling with kernel size k to the detached attention map p_g^t, followed by bilinear upsampling to the original resolution, and thresholding it into a binary foreground mask $M_g^t \in \{0,1\}^{H\times W}$ by retaining the locations whose envelope value is at least a fraction τ of its maximum. Among the 8-connected components of M_g^t, those below a minimum area $a_{\min}$ are discarded, leaving the valid candidate regions $\mathcal{C}_g^t$ of this identity. All masks are built from detached attention, so that gradients reach the model only through the p_g^t that appears in the penalty below.

Among these candidates, we select an anchor $\hat{M}_g^t \in \{0,1\}^{H\times W}$ as the dominant region of the g-th identity at frame t. When the anchor of the previous frame, $\hat{M}_g^{t-1}$, is available, we select the component that overlaps most with it, so that the same region is tracked over time; otherwise, we select the component with the largest attention mass:

$$\hat{M}_g^t = \begin{cases} \arg\max_{\mathcal{C}\in\mathcal{C}_g^t} \left|\mathcal{C} \cap \hat{M}_g^{t-1}\right|, & \hat{M}_g^{t-1} \neq \mathbf{0}, \\ \arg\max_{\mathcal{C}\in\mathcal{C}_g^t} \sum_{(h,w)\in\mathcal{C}} p_g^t(h,w), & \hat{M}_g^{t-1} = \mathbf{0}. \end{cases} \tag{2}$$

Here, $|\cdot|$ denotes the number of pixels, and $\hat{M}_g^{t-1} = \mathbf{0}$ indicates that no previous anchor is available. We also fall back to attention-mass selection when the maximum temporal overlap is below a threshold τ_{overlap}, and set $\hat{M}_g^t = \mathbf{0}$ if no valid component remains.

With the anchor determined, $M_g^t(1-\hat{M}_g^t)$ marks the locations that lie inside the foreground mask but outside the anchor, i.e., the competing disconnected activations of the g-th identity. Let $\mathcal{A}_t$ denote the set of identities with at least one valid component in frame t. We compute the Tracking Loss for each identity in each frame, normalize it by the corresponding foreground-mask area, and then average over active identities and frames:

$$\mathcal{L}_{\text{track}}^{(g,t)} = \frac{\sum_{h,w} p_g^t M_g^t (1-\hat{M}_g^t)}{\sum_{h,w} M_g^t + \varepsilon}, \qquad \mathcal{L}_{\text{track}} = \frac{1}{T}\sum_t \frac{1}{|\mathcal{A}_t|} \sum_{g\in\mathcal{A}_t} \mathcal{L}_{\text{track}}^{(g,t)}, \tag{3}$$

where $\varepsilon > 0$ is a small constant for numerical stability. Because the penalty lives entirely inside the foreground mask, background attention is untouched and only the rival regions of the same identity are pushed down, leaving each identity with a single, temporally coherent region and therefore far fewer duplicated instances.

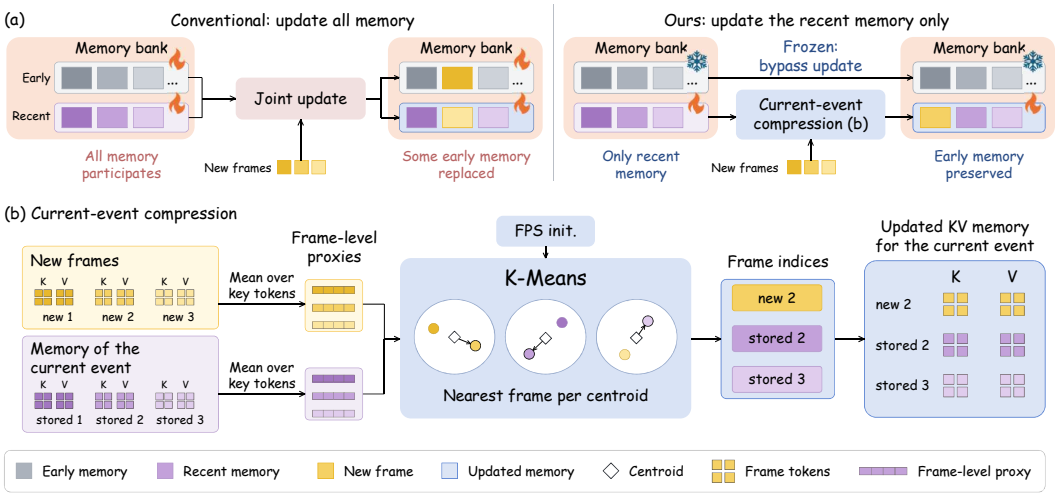


Figure 5: **Event-wise memory compression in MemID.** Only the memory of the current event is updated, while memories of completed events are frozen and retained as persistent historical context.

Training Objective. We incorporate the ID-aware Attention Losses into the DMD distillation objective, $\mathcal{L}_{\text{DMD}}$, where the Exclusive Loss and Tracking Loss are jointly weighted to constrain the attention maps. The overall training objective is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{DMD}} + w_{\text{excl}}\mathcal{L}_{\text{excl}} + w_{\text{track}}\mathcal{L}_{\text{track}}, \tag{4}$$

where w_{excl} and w_{track} balance the contributions of each loss.

3.2 ID-PRESERVING MEMORY

To extract relevant information from long-term history without incurring the prohibitive quadratic cost of full-context attention as the history grows, MemID factorizes context construction into two stages: *Event-wise Memory Compression* and *Text-Video Context Retrieval*. During Event-wise Memory Compression, MemID maintains a memory bank by continuously updating the memory of the ongoing event, while storing the memories of completed events as persistent historical memory. Subsequently, during Text-Video Context Retrieval, MemID retrieves prompt-relevant context from the memory bank based on text-video similarity.

Event-wise Memory Compression. Conventional memory mechanisms update the memory using a fixed bank strategy, where they continuously replace historical memories with newly generated content, which may discard earlier reliable anchors and cause the memory itself to drift with AR generation. To mitigate this, we adopt the Event-wise Memory Compression illustrated in Fig. 5(a). Frames generated under the same prompt belong to the same event. Only the memory of the current event is updated, while memories from completed events are frozen. Within each event, representative frames are selected to maintain a compact and stable memory. As shown in Fig. 5(b), we first construct a candidate frame set for the current event by combining its previously stored memory frames with the newly generated frames. For each candidate frame I_j, we compute a frame-level proxy p_j from its token-level keys, and cluster the resulting proxy set P into N centroids:

$$p_j = \overline{\pi(I_j)}, \qquad C = \text{K-Means}(P, N), \tag{5}$$

where $\pi(\cdot)$ extracts the token-level keys and the overline denotes token averaging. Here, P denotes the set of proxies for all candidate frames and $C = \{c^n\}$ denotes the resulting cluster centroids. For each centroid c^n, we select the candidate frame whose proxy is closest to it, and use the original token-level keys and values of the selected frames to update the memory of the current event. This nearest-frame selection avoids directly storing averaged centroids, which may lose fine-grained identity details. We initialize the K-Means centroids using Farthest Point Sampling (FPS) for improved stability. Once the prompt switches, the resulting event memory is frozen and retained throughout subsequent generation.

Text-Video Context Retrieval. Given the accumulated memories of historical events, we retrieve relevant context for current generation using text-video similarity. This allows the model to access identity evidence from distant events without feeding the entire video history into the attention module. Formally, let K_u denote the token-level keys of the u-th memory entry and let t denote the current text prompt. We compute the text-video similarity ϕ_u as

$$\phi_u = \sum_{s=1}^{S} \left\langle t_s, \overline{K_u} \right\rangle, \tag{6}$$

where t_s denotes the key of the s-th text token, S is the number of text tokens, $\overline{K_u}$ denotes the token-averaged visual key of the u-th memory entry, and $< \cdot, \cdot >$ denotes the inner product.

We select the top-k memory entries according to ϕ_u. Let $\hat{u}$ denote their indices and V denote the corresponding values in the historical memory bank. We then feed $K_{\hat{u}}$ and $V_{\hat{u}}$ into the attention module as additional context tokens. In this way, MemID preserves historical coverage across events while exposing only a compact, prompt-adaptive subset of the memory to each generation step.

4 Experiments

MemID is trained for 2,800 optimization steps with a batch size of 8 using the AdamW optimizer (Loshchilov & Hutter, 2017). We use the switch-prompt training dataset introduced by LongLive (Yang et al., 2025) and employ Qwen2.5-32B-Instruct (Qwen Team, 2024) to identify entity IDs in the training prompts. The memory size for each event and the final retrieved memory size are set to 3 frames. We employ the narrative multi-prompt benchmark introduced by MemFlow (Ji et al., 2025), generating 60-second videos with a new prompt provided every 10 seconds to introduce complex character entrances and interactions. For evaluation, we adopt the standard VBench-Long (Huang et al., 2024) metrics. Specifically, we use the VBench-Long Quality Score to measure overall quality, define the Consistency Score as the average of Subject Consistency and Background Consistency, and use the Aesthetic Score to assess visual quality. In addition, we introduce an *ID Consistency Score* and an *ID Counting Score* to evaluate identity preservation. The *ID Consistency Score* measures whether prompt-specified entities are consistently associated with their corresponding instances by comparing appearance features against a reference established upon the first successful localization using DINO v2 features (Oquab et al., 2023). The *ID Counting Score* measures whether the number of generated instances matches the ground-truth count. Detailed definitions of both metrics are provided in the Appendix.

4.1 Comparison with Other Methods

We evaluate our model, ***MemID***, against six state-of-the-art multi-prompt streaming video generation models: Self-Forcing (Huang et al., 2025b), Context-Forcing (Chen et al., 2026), Infinity-RoPE (Yesiltepe et al., 2026), LongLive (Yang et al., 2025), MemFlow (Ji et al., 2025), and IAMFlow (Liu et al., 2026). Quantitatively, as shown in Tab. 1, MemID achieves the best ID Consistency Score and ID Counting Score among the compared methods. Notably, MemID maintains its lead in ID Consistency Score across subsequent narrative events, indicating that MemID effectively alleviates identity drift across distant narrative events. MemID also achieves the best Consistency Score and the second-best Aesthetic Score, while maintaining a Quality Score comparable to the baselines. As illustrated in Fig. 6, MemID achieves superior character preservation and structural coherence compared with existing state-of-the-art methods. Most baselines suffer from ID duplication, largely due to attention leakage in cross-attention, which severely compromises semantic consistency. In contrast, MemID consistently preserves the correct character identities, benefiting from the explicitly constrained attention maps. Moreover, MemFlow (Ji et al., 2025), LongLive (Yang et al., 2025), IAMFlow (Liu et al., 2026), and Context-Forcing (Chen et al., 2026) exhibit noticeable ID drift as generation progresses, while Self-Forcing (Huang et al., 2025b) suffers from rapid

Table 1: **Comparison with other baseline models on the narrative multi-prompt benchmark.**

Method	Quality Score ↑	Consistency Score ↑	Aesthetic Score ↑	ID Counting Score ↑	ID Consistency Score ↑						
					Avg.	0–10s	10–20s	20–30s	30–40s	40–50s	50–60s
Self-Forcing	77.71	96.88	57.37	66.66	41.96	85.99	59.74	42.89	29.33	21.08	12.70
IAMFlow	79.05	<u>97.35</u>	59.42	71.44	58.46	87.68	61.56	54.69	49.95	45.93	50.94
Infinity-RoPE	<u>80.26</u>	97.32	60.28	70.88	62.07	**90.77**	<u>69.56</u>	<u>61.96</u>	54.38	49.50	46.27
Context-Forcing	**80.68**	97.00	**61.25**	67.83	57.32	87.79	68.75	55.09	49.64	44.42	38.24
LongLive	79.31	97.13	60.44	71.60	<u>62.45</u>	88.17	67.70	58.32	<u>55.56</u>	<u>52.92</u>	<u>52.01</u>
MemFlow	78.89	97.19	59.02	<u>71.70</u>	58.02	88.18	63.02	55.46	49.10	46.00	46.34
MemID (Ours)	78.96	**97.75**	<u>60.76</u>	**74.21**	**65.34**	<u>89.67</u>	**69.93**	**63.48**	**59.58**	**55.79**	**53.57**



Figure 6: **Qualitative comparison with other baseline models.** *MemID* maintains more consistent character identities across long narratives.

Table 2: **Ablation study on the narrative multi-prompt benchmark.** "w/o Attn. Losses" means that the model is trained using only DMD Loss, and "w/o ID. Mem." means that the model replaces the ID-preserving Memory with the MemFlow-style memory mechanism.

Method	Quality Score ↑	Consistency Score ↑	Aesthetic Score ↑	ID Counting Score ↑	ID Consistency Score ↑						
					Avg.	0–10s	10–20s	20–30s	30–40s	40–50s	50–60s
w/o Attn. Losses	**79.33**	96.49	58.86	68.74	53.02	86.48	67.03	51.84	42.17	37.40	33.20
w/o ID. Mem.	78.48	97.04	59.04	71.00	55.38	86.40	62.33	51.88	47.30	44.47	39.90
Full Model	78.96	**97.75**	**60.76**	**74.21**	**65.34**	**89.67**	**69.93**	**63.48**	**59.58**	**55.79**	**53.57**

error accumulation over time. By contrast, MemID maintains high visual fidelity and stable character identities throughout the entire 60-second generation, which we attribute to our ID-preserving Memory.

4.2 Ablation Study

Main Ablation Study. We ablate the two core components of MemID, the *ID-preserving Memory* and the *ID-aware Attention Losses*, on the narrative multi-prompt benchmark (Fig. 7, Table 2). Replacing our memory with the continuously refreshed update of MemFlow (Ji et al., 2025), which keeps no frozen anchor for completed events, consistently degrades long-range ID consistency and invites *temporal memory drift*, *e.g.*, the apprentice's uniform shifts in color across events. Removing the attention losses fragments identity-specific attention into disconnected regions and induces *ID duplication*, such as the emergence of a second apprentice (first two rows of Fig. 7). Persistent historical anchors and explicit attention regularization are thus both indispensable for maintaining identity over long narratives. We ablate the remaining components below.

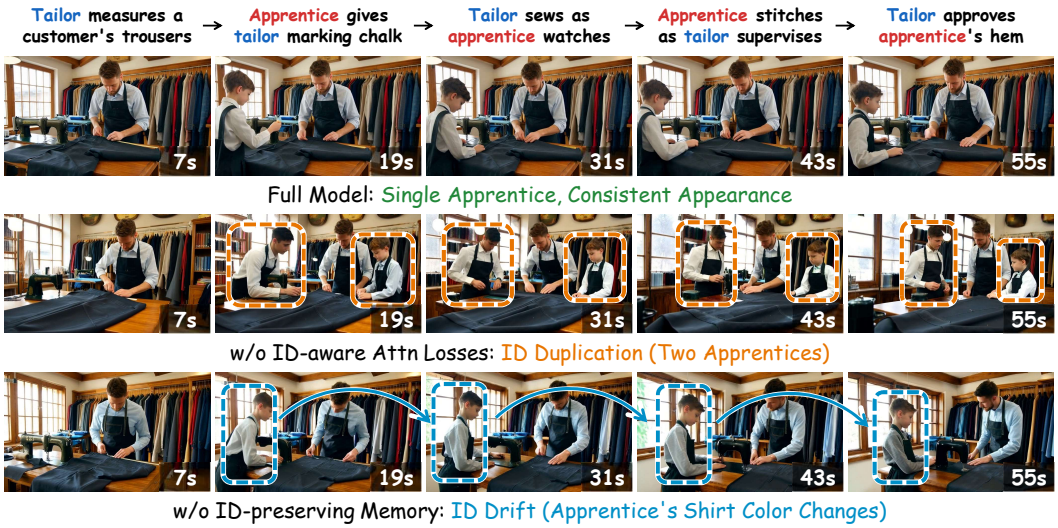


Figure 7: **Qualitative evaluation of the ablation study.** Adding the ID-aware Attention Losses can mitigate ID duplication, and adding ID-preserving Memory can mitigate ID drift.

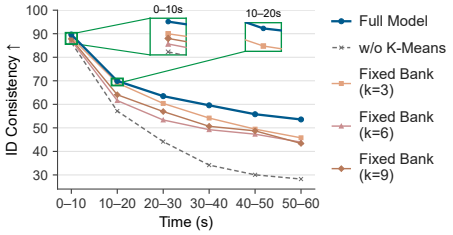


Figure 8: **Ablation study on Event-wise Memory Compression.** Full model outperforms the variants in most events.

Method	Consistency Score ↑	ID Consistency ↑	ID Counting ↑
DMD Loss	96.49	53.02	68.74
+ Track. Loss	97.21 (+0.72)	59.21 (+6.19)	72.92 (+4.18)
+ Excl. Loss	96.95 (+0.46)	57.50 (+4.48)	73.20 (+4.46)
+ Attn. Losses	**97.75** (+1.26)	**65.34** (+12.32)	**74.21** (+5.47)

Table 3: **Ablation study on the components of ID-aware Attention Losses.** "+ Track. Loss" and "+ Excl. Loss" indicate adding the Tracking Loss and Exclusive Loss in training individually, respectively, while "+ Attn. Losses" denotes using both losses jointly.

Event-wise Memory Compression. To verify the event-wise memory compression in our ID-preserving memory, as shown in Fig. 8, we first replace K-Means clustering with a text-video-similarity-based strategy, which leads to a significant decrease in ID Consistency. In addition, we replace our event-wise bank update strategy with a fixed-size memory bank. The fixed bank keeps a bank of fixed size $k \in \{3, 6, 9\}$ that is continuously overwritten as generation proceeds, whereas our Full Model stores compressed frames per event and freezes them once the event completes. From Fig. 8, simply increasing the bank capacity without event-wise consideration does not bring consistent performance gains.

Details of the Attention Losses. The proposed Attention Losses contain the Tracking Loss and the Exclusive Loss. As shown in Table 3, the Tracking Loss improves ID Consistency Score by 6.19 points and ID Counting Score by 4.18 points, while also yielding a 0.72-point gain in Consistency Score. The Exclusive Loss provides consistent improvements across all three metrics, including gains of 4.48 points in ID Consistency Score and 4.46 points in ID Counting Score. Notably, combining both losses increases ID Consistency Score by 12.32 points and ID Counting Score by 5.47 points over the baseline, demonstrating their complementary effects on identity preservation.

5 Conclusion

We presented MemID, an identity-aware AR video diffusion framework for preserving long-term character consistency in streaming multi-prompt narrative video generation. MemID addresses two complementary sources of identity degradation: temporal memory drift and spatial identity interference. Its ID-preserving Memory consolidates and freezes event-specific memories to maintain persistent historical identity anchors, while its ID-aware Attention Losses suppress inter-ID attention overlap and intra-ID attention fragmentation to mitigate ID fusion and duplication. On a narrative multi-prompt benchmark, MemID substantially outperforms state-of-the-art baselines in ID Consistency Score and ID Counting Score while maintaining competitive Aesthetic Score. Ablation studies further confirm that both components contribute significantly to long-range identity preservation.

AI USE STATEMENT

In this work, we used generative AI tools for three purposes: (1) constructing synthetic datasets, (2) simplifying prompts used for result presentation to make them clearer and more concise, and (3) improving the readability, grammar, and clarity of the manuscript. All AI-assisted outputs were manually reviewed, including verification of language edits and validation of AI-generated synthetic data before their use in experiments. We take responsibility for the final content of this work, including text, claims, data, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the key methodological and experimental details in the main text, including the training setup, memory configuration, and evaluation protocol. Additional implementation details, including the complete training configuration, loss schedules, hyperparameters, and the full definitions of *ID Consistency Score* and *ID Counting Score*, are provided in the supplementary material. We also provide detailed descriptions of the data construction and processing procedures in the supplementary material. Code and data will be made publicly available upon acceptance.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. *CVPR*, 2025.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5331–5341. IEEE, 2024.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Lingjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, Yingnian Wu, and Lijuan Wang. Slowfast-vgen: Slow-fast learning for action-conditioned long video generation. *arXiv*, 2024.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

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

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

Team HunyuanWorld. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint*, 2025.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8079–8088. IEEE, 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, 2025. URL https://arxiv.org/abs/2512.14699.

Weinan Jia, Yuning Lu, Mengqi Huang, Hualiang Wang, Binyuan Huang, Nan Chen, Mu Liu, Jidong Jiang, and Zhendong Mao. Moga: Mixture-of-groups attention for end-to-end long video generation. *arXiv preprint arXiv:2510.18692*, 2025.

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

Subin Kim, Seoung Wug Oh, Jui-Hsien Wang, Joon-Young Lee, and Jinwoo Shin. Tuning-free multi-event long video generation via synchronized coupled sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6418–6429, 2025.

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7667–7677. IEEE, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023.

Zhiyuan Li, Chi-Man Pun, Chen Fang, Jue Wang, and Xiaodong Cun. Personalive! expressive portrait image animation for live streaming. *arXiv preprint arXiv:2512.11253*, 2025b.

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*, 2026.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

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

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=X9Fga52OOv.

Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*, 2025.

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.

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

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahan Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, et al. Evoworld: Evolving panoramic world generation with explicit 3d memory. *arXiv preprint arXiv:2510.01183*, 2025.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7418–7427. IEEE, 2023.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

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. *arXiv preprint arXiv:2509.22622*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

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*, 2025b.

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21685–21695, 2025c.

A Identity Attention Map

To obtain the attention map for each identity in the training prompts, we first use an LLM to identify the identities mentioned in each prompt and then average the cross-attention maps of the tokens corresponding to each identity. The system prompt for the LLM is as follows.

```
You will receive a list of scene descriptions in order.

All scenes belong to the same continuous narrative.

IMPORTANT: Focus ONLY on FOREGROUND entities.

Foreground entities are the main subjects that actively participate
in the narrative (e.g., main characters, people, animals, or key
objects being directly interacted with).

Do NOT treat atmosphere, scenery, lighting, weather, locations, or
background props as entities, even if they are described in detail.

CRITICAL CONSTRAINT ON ENTITY NAMES:

Entity names MUST be chosen from words that explicitly appear in the
input text.

You MAY extract a single word or a short contiguous phrase from a
longer description (for example, extracting 'detective' from 'a weary
detective in his 50s').

You MUST NOT replace an entity with a synonym, paraphrase, or any
word that does NOT appear in the input text.

Do NOT normalize or translate entity names.

Your task is to:
    1. Read and understand all scenes as a whole.
    2. Identify ONLY the core foreground entities that appear
       repeatedly or play a key role in driving the narrative.
    3. Select entity names by copying words or short phrases directly
       from the input text.
    4. Prefer concise entity names, but only if they are exact words
       present in the text.
    5. Ignore entities that appear only once or are not narratively
       important.
```

```
    6. Keep the number of entities minimal and focused (typically 2
       to 5).
Output ONLY a valid JSON object exactly like:
      {"ref_id":  ["entity_1", "entity_2"]}
Do not include any explanations, comments, or additional text.
```

For each identity i, we first average the cross-attention maps over all text tokens associated with the identity, $\mathcal{T}_i$, as well as over all attention heads and participating DiT blocks. We then apply min-max normalization to obtain the identity-specific attention map p_i:

$$A_i = \frac{1}{N_{\text{block}} N_{\text{head}} |\mathcal{T}_i|} \sum_{b=1}^{N_{\text{block}}} \sum_{h=1}^{N_{\text{head}}} \sum_{u \in \mathcal{T}_i} \text{Attn}_{b,h,u}, \qquad p_i = \frac{A_i - \min A_i}{\max A_i - \min A_i + \varepsilon}, \tag{7}$$

where N_{head} is the number of attention heads, N_{block} is the number of participating DiT blocks, and ε is a small constant introduced for numerical stability. The minimum and maximum are computed over all spatial locations.

B Implementation Details

We initialize MemID from the LongLive Stage 1 checkpoint and train it for 2,800 optimization steps with a batch size of 8 using the AdamW optimizer (Loshchilov & Hutter, 2017). MemID supports streaming video generation at over 16 frames per second. The Exclusive Loss scale w_{excl} is warmed up from 0 to 1.0 over iterations 400–600 and annealed to 0 over iterations 800–1000, and the Tracking Loss scale w_{track} is warmed up from 0 to 0.25 over iterations 400–600 and annealed to 0 over iterations 800–1000. Our ID-aware Attention Losses are applied only to the 13th through 17th DiT blocks. We compute the ID-aware Attention Losses only at the first two denoising steps. Qwen2.5-32B-Instruct (Qwen Team, 2024) is used to identify entity IDs from the training prompts.

C ID Consistency Score

We measure identity consistency by tracking the entities associated with the IDs specified in the text prompt throughout the generated video. For each prompt, we first use DeepSeek (Guo et al., 2025) to extract the entity IDs. These IDs are then used as textual queries for GroundingDINO (Liu et al., 2024) to localize the corresponding entities in sampled video frames. For each frame, GroundingDINO produces a set of candidate bounding boxes together with their detection scores.

For each entity ID, we sequentially scan the generated video frames until the entity is successfully localized. Specifically, the first frame in which the maximum detection score among all candidate boxes associated with the ID exceeds 0.3 is designated as the anchor frame. We select the candidate box with the highest detection score in this frame, enlarge it by 10%, and crop the corresponding region to serve as the anchor for that entity.

Once the anchor is established, we localize the same entity in subsequent frames. For each frame, GroundingDINO (Liu et al., 2024) generates candidate detection boxes, which are retained only if their detection scores exceed 0.3 and their DINO (Oquab et al., 2023) feature cosine similarity to the corresponding anchor exceeds 0.5. If no candidate satisfies both criteria, the entity is considered unsuccessfully localized in that frame. Otherwise, we select the candidate box with the highest DINO cosine similarity to the anchor.

To prevent the same visual instance from being assigned to multiple entity IDs, entities are matched in the order of their first appearance in the narrative, with earlier-appearing IDs given priority. For each candidate detection, we obtain its segmentation mask using SAM (Kirillov et al., 2023). A candidate is excluded if its mask has an IoU greater than 0.5 with any mask already assigned to another entity; otherwise, it is selected as the matched instance.

For each entity ID i, we extract the DINO v2 (Oquab et al., 2023) feature from its anchor as a fixed reference feature, denoted by $\mathbf{f}_i^{\text{ref}}$. This feature serves as the identity reference for all subsequent successful localizations of the same entity. For entity i at sampled frame t, we define the frame-level

identity consistency score as

$$c_{i,t} = \begin{cases} 0, & \text{if the entity is not successfully localized,} \\ 1, & \text{for its first successful localization,} \\ \max\left(0, \cos(\mathbf{f}_{i,t}, \mathbf{f}_i^{\text{ref}})\right), & \text{otherwise,} \end{cases}$$

where $\mathbf{f}_{i,t}$ denotes the DINO feature extracted from the localized entity at frame t. The first successful localization is assigned a score of 1 by definition because its feature is used to establish the reference identity. This formulation jointly accounts for both localization failures and visual appearance deviations from the entity's initial reference.

For each event e, we evaluate only the entities expected to appear in that event. Let $\mathcal{I}_e$ denote the set of active entity IDs and $\mathcal{F}_e$ the set of sampled frames associated with event e. For each active entity i, its event-level consistency score is computed by averaging the frame-level scores over all sampled frames:

$$C_{e,i} = \frac{1}{|\mathcal{F}_e|} \sum_{t \in \mathcal{F}_e} c_{i,t},$$

where unsuccessful localizations contribute a score of 0.

Finally, the ID Consistency Score for event e is computed by uniformly averaging the event-level consistency scores over all active entities:

$$C_e = \frac{1}{|\mathcal{I}_e|} \sum_{i \in \mathcal{I}_e} C_{e,i}.$$

D ID Counting Score

ID Counting Score evaluates whether the number of generated instances for each prompt-specified entity matches the expected count. We derive the ground-truth counts from the scene descriptions using an LLM (Qwen Team, 2024). For each entity in each scene, the count is specified as zero if the entity is not required, a positive integer if an explicit count is available, and unknown otherwise. Entities with unknown counts are excluded from counting evaluation but remain valid for *ID Consistency Score*.

For each sampled frame, we detect candidate instances and obtain their masks using SAM (Kirillov et al., 2023). To avoid duplicated detections, detections within each entity are clustered based on mask overlap, while cross-entity duplicates are removed according to mask overlap and detection confidence. For an entity with ground-truth count n, the frame-level score is

$$s_{\text{count}} = \begin{cases} 1, & N_{\text{det}} = n, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where N_{det} denotes the number of retained instances. The final score is obtained by averaging the frame-level scores over entities, events, and videos.

E Visualization of Attention Maps

We visualize the attention maps of MemID and several baseline models to further validate the effectiveness of our ID-aware Attention Losses. As shown in Figure 9, MemID better preserves the appearance of both the writer and the cat, whereas the baseline models often suffer from ID duplication and identity fusion. This phenomenon is also clearly reflected in the attention maps. Compared with the overlapping and fragmented attention patterns produced by the baseline models, MemID generates more distinct and well-separated attention maps, with each map concentrated within a compact spatial region corresponding to its associated subject.

F Visualization of Memory Mechanism

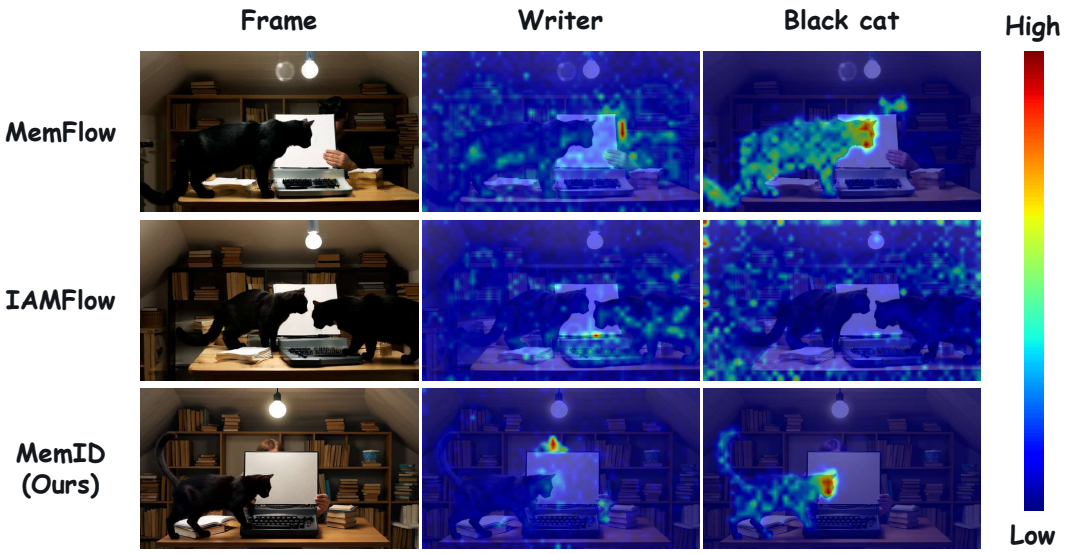


Figure 9: **Attention maps for a prompt containing a writer and a black cat.** MemID produces better-separated and more compact identity attention maps, leading to better identity consistency.

To further quantitatively demonstrate the effectiveness of our ID-preserving Memory in preserving early frames, we compute the mean frame index of the context used by the attention module during generation. Specifically, the current index denotes the index of the frame being generated, while the mean context index represents the average frame index of the frames retained in its context. For the same current index, a lower mean context index indicates that the context contains earlier frames. The $y = x$ line serves as a reference where the mean context index is equal to the current frame index; thus, a curve farther below this line indicates that the retained context comes from relatively earlier frames. As shown in Figure 10, MemID consistently exhibits a lower mean context index than MemFlow and IAMFlow throughout generation, with a more pronounced difference in later frames. This demonstrates that MemID better retains early-frame information in the context, thereby preserving early character anchors over long-term generation.

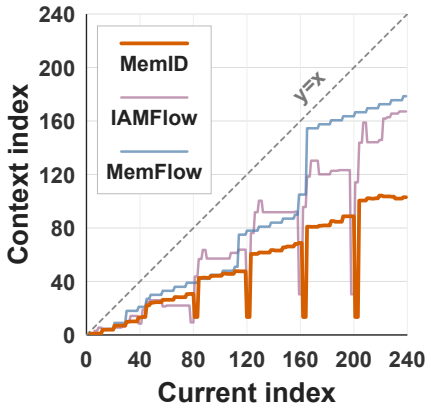


Figure 10: **Mean context frame index during generation.** Lower mean indices indicate that earlier frames are retained in the context.

G More Qualitative Results

We provide more qualitative results generated by our MemID as shown in Figures 11, 12, 13 and 14.

H Limitation

MemID remains less robust to collective or relational expressions that refer to multiple individual identities, such as “a couple.” The model may interpret such expressions as a single semantic entity rather than two distinct characters, leading to incorrect identity association or generation. Improving the decomposition and grounding of such multi-entity expressions is an important direction for future work.



Figure 11: More qualitative results from MemID.



0-10s: The flour-dusted baker kneads dough.



10-20s: A little girl enters and peers at the baker's pastries.



20-30s: The baker asks the little girl what she wants.



30-40s: The baker gives the girl a croissant and play dough.



40-50s: The girl eats and plays while the baker resumes kneading.



50-60s: The baker works and smiles at the little girl.

Figure 12: More qualitative results from MemID.



0-10s: The father chops wood beside the snowy cabin.



10-20s: The son carries a log to help his father.



20-30s: The father stops chopping and accepts his son's log.



30-40s: The father teaches his son to stack wood safely.



40-50s: The son throws a snowball at his father after stacking.



50-60s: The father and son have a playful snowball fight.

Figure 13: More qualitative results from MemID.

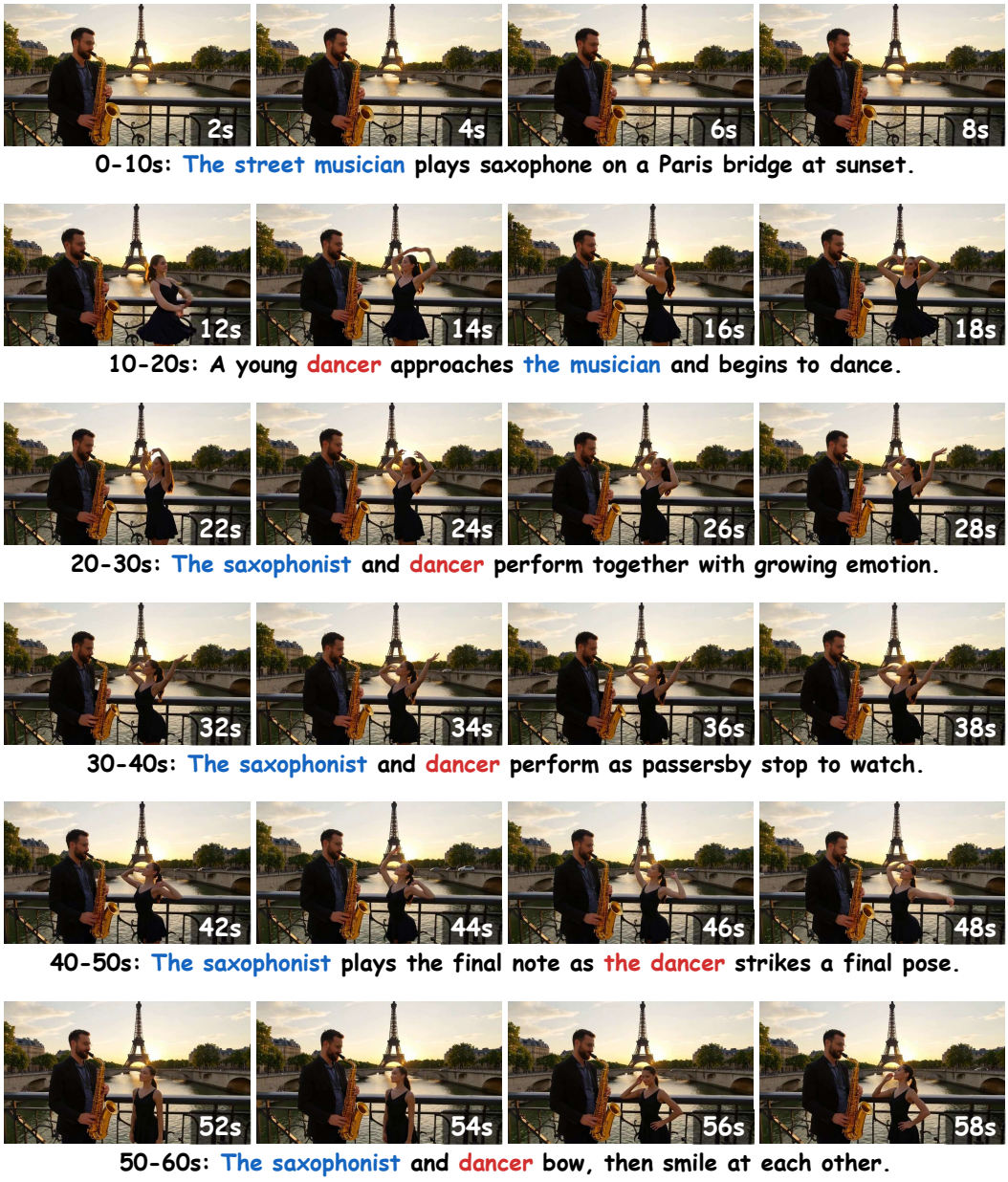


Figure 14: More qualitative results from MemID.