MEMENTO: RECONSTRUCT TO REMEMBER FOR CONSISTENT LONG VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form video generation requires recurring subjects to remain consistent across shots, viewpoints, and scene transitions. Memory-conditioned autoregressive methods improve scalability through shot-by-shot generation. However, they primarily optimize memory conditioning or selection mechanisms, without explicitly supervising the utilization of identity-critical evidence in historical memory, allowing subject appearance to drift over time. We propose **Memento**, a subject-reconstruction-guided framework that treats subject preservation as an explicit identity grounding problem, introducing memory-based subject reconstruction to encourage the model to preserve and retrieve identity-critical appearance cues. Memento jointly trains next-shot generation and subject reconstruction, recovering target appearances from historical memory and subject descriptions. A dual-query memory mechanism separates these historical cues: story-conditioned queries retrieve long-range subject evidence, while shot-conditioned queries select short-range contextual references for coherent continuation. A subject-aware cinematic data pipeline aligns story, shot, and reconstruction captions through consistent subject descriptions, providing unambiguous supervision for identity grounding. Experiments demonstrate state-of-the-art subject consistency, particularly across scene transitions, together with strong cross-shot coherence and competitive visual quality.

1 INTRODUCTION

Recent diffusion-based video generation models have achieved remarkable progress in visual fidelity, motion realism, and text alignment Ho et al. (2020); Peebles & Xie (2023); Blattmann et al. (2023); Rombach et al. (2022); Wan et al. (2025); Kong et al. (2024); Yang et al. (2024); Google DeepMind (2026); Kuaishou (2026); OpenAI (2026); Yu et al. (2025); Seedance et al. (2026). Despite these advances, generating long-form and story-driven videos remains challenging. A long video is not merely an extended sequence of visually plausible frames, it requires recurring subjects to maintain their identities across multiple shots, viewpoints, poses, motions, and scene transitions. However, subtle changes in face, clothing, body shape, or object appearance can accumulate over time, causing subject drift and breaking narrative coherence.

Existing long video generation methods typically address this challenge through temporal decomposition. Specifically, storyboard-based approaches Zhou et al. (2024); Xie et al. (2024); Zhao et al. (2024); Zheng et al. (2024); Hu et al. (2024); He et al. (2025); Xiao et al. (2026); Zhang et al. (2025b); Long et al. (2024); Huang et al. (2024a) synthesize sparse keyframes before animating them into clips, while joint multi-shot methods Guo et al. (2025); Meng et al. (2025); Wu et al. (2025); Qi et al. (2025); Kara et al. (2025); Cai et al. (2025); Jia et al. (2025) generate multiple shots within a shared diffusion context. Although these strategies improve cross-shot coordination, they either leave individual clips largely independent or remain constrained by context length and escalating long-range attention costs. Alternatively, memory-conditioned autoregressive methods Zhang et al. (2025a); An et al. (2025) offer a scalable alternative by generating videos shot by shot with compact historical memory. However, their memory selection mechanisms are usually driven by generic relevance, visual saliency, or context compression, rather than explicit identity preservation. As generation proceeds, identity-critical evidence of recurring subjects may be gradually diluted.

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.

We contend that subject consistency in long video generation is fundamentally an identity grounding problem. Existing memory-conditioned methods typically use historical frames as conditions for next-shot generation, where the training signal primarily rewards plausible local continuation. While such supervision can improve temporal coherence in the short term, it does not explicitly guarantee that the memory retains fine-grained identity evidence for maintaining subjects. Consequently, the model may learn to exploit memory for scene layouts or visual continuity, while failing to preserve the critical appearance cues required to maintain the same subject across distant shots.

Our key intuition is that subject consistency should be explicitly learned rather than expected to emerge as a by-product of optimizing shot-level generation quality. Even with access to historical memory, the model may underutilize identity-critical cues, resulting in gradual appearance drift. To address this limitation, we introduce Subject-Anchored Multi-Task Training, which jointly optimizes shot generation and memory-based subject reconstruction. The reconstruction task recovers target subject appearances using only historical memory and subject descriptions, without direct visual prompts. This encourages the model to preserve and retrieve persistent identity cues, including facial features, clothing textures, and body structure, thereby improving subject consistency across shots.

This reconstruction-guided perspective further suggests that a single entangled memory is suboptimal for long video generation. Subject reconstruction and next-shot generation depend on different forms of historical information. Specifically, reconstruction benefits from long-range identity evidence that should remain stable across scene changes and distant shots. In contrast, next-shot generation relies more heavily on short-range contextual cues, such as recent layout, motion tendency, and camera continuity. Consequently, when both objectives share the same memory selection policy, visually salient recent context can easily dominate the memory, while sparse but identity-critical evidence may be suppressed. This motivates a disentangled memory design that separately retrieves long-context subject evidence for identity grounding and short-context visual references for local shot generation.

Based on these ideas, we propose **Memento**, which combines reconstruction-guided training, dual-path memory, and subject-aware data supervision for autoregressive long video generation under a fixed memory budget. Our contributions are summarized as follows:

- We propose **Memento** with Subject-Anchored Multi-Task Training, which jointly optimizes next-shot generation and memory-based subject reconstruction using a shared generator. The reconstruction objective provides explicit subject-level supervision to the generation pathway, improving its ability to recover recurring subject appearances from long-range memory without introducing inference-time overhead.
- We design a dual-path memory mechanism that accommodates the distinct historical requirements of identity recovery and shot continuation. Story-conditioned queries retrieve long-range subject evidence, whereas shot-conditioned queries select short-range contextual cues for coherent next-shot generation, enabling scalable autoregressive synthesis under a fixed memory budget.
- We construct a subject-aware cinematic data curation pipeline that associates recurring subjects across global story, per-shot, and reconstruction captions using consistent, pronoun-free descriptions, providing unambiguous supervision for joint generation and reconstruction.

2 Related Work

2.1 Single-Shot Video Generation

Recent advances in diffusion models Ho et al. (2020) and Diffusion Transformers (DiTs) Peebles & Xie (2023) have substantially improved video generation. Large-scale text-to-video (T2V) and image-to-video (I2V) models Singer et al. (2022); Ho et al. (2022); Blattmann et al. (2023); Wan et al. (2025); Kong et al. (2024); Yang et al. (2024); Google DeepMind (2026); Kuaishou (2026); OpenAI (2026); Yu et al. (2025); Seedance et al. (2026) can now synthesize short, single-shot clips with high visual fidelity and natural motion. However, these models are primarily designed for temporally bounded single scenes and lack explicit mechanisms for maintaining consistency across multiple shots. Consequently, they remain insufficient for long-form, story-driven video generation,

where recurring subjects, scenes, and narrative logic must remain coherent over extended sequences. This limitation motivates the development of multi-shot video generation approaches.

2.2 Multi-Shot Long Video Generation

Multi-shot video generation aims to produce long-form narratives by decomposing video synthesis into shot-level generation. Existing methods mainly follow three paradigms.

Storyboard-based generation. Storyboard-based methods Xie et al. (2024); Zhao et al. (2024); Zheng et al. (2024); Hu et al. (2024); Zhou et al. (2024); He et al. (2025); Xiao et al. (2026); Zhang et al. (2025b); Long et al. (2024); Huang et al. (2024a) first generate sparse keyframes and then animate them with pretrained image-to-video models. Although efficient, they enforce consistency mostly at the keyframe level, leaving individual clips weakly coupled and often causing rigid transitions, detail drift, and limited narrative coherence.

Joint multi-shot generation. Joint generation methods synthesize multiple shots within a single diffusion forward pass. Representative works such as LCT Guo et al. (2025) and HoloCine Meng et al. (2025) introduce cross-shot attention mechanisms to improve global interaction and shot-level alignment, with later methods further improving efficiency and controllability Wu et al. (2025); Qi et al. (2025); Kara et al. (2025); Cai et al. (2025); Jia et al. (2025). However, these methods remain bounded by the context window, and their memory and computation costs increase rapidly as the number of shots grows, limiting open-ended generation.

Memory-conditioned autoregressive generation. Autoregressive methods scale better by generating videos shot by shot while conditioning on a compact memory of previous content. StoryMem Zhang et al. (2025a) maintains selected keyframes using semantic and aesthetic criteria, while OneStory An et al. (2025) compresses historical context for next-shot generation. These designs improve scalability, but their memory selection is primarily based on generic relevance or short-term context compression, rather than explicit subject preservation. As a result, identity-critical cues may be diluted or overwritten over long generation horizons.

In contrast, **Memento** explicitly formulates long-form consistency as subject grounding in memory. It uses a dual-query memory mechanism to separate long-context subject evidence from short-context generation cues, and jointly optimizes next-shot generation with memory-based subject reconstruction. This encourages the memory to retain identity-critical information while preserving local visual continuity, enabling more stable long video generation.

3 Method

Memento follows the principle that consistent long video generation requires both persistent subject evidence and recent contextual cues. To this end, we generate videos autoregressively at the shot level using a compact historical memory bank. Crucially, we introduce subject reconstruction as an explicit supervision signal, which forces the model to actively retrieve and utilize identity-critical features stored in the memory. This design allows the model to preserve recurring identities without attending to all previous frames, while still using recent visual context for local shot continuity.

Given a global story caption C_{story} and a sequence of per-shot captions $\{C^t_{\text{shot}}\}_{t=1}^N$, Memento generates a multi-shot video $V = S_1, \ldots, S_N$. The first shot is generated by standard text-to-video synthesis. For each subsequent shot $t \geq 2$, the model performs memory-conditioned generation using C_{story}, C^t_{shot}, and a dynamically updated memory bank M_t. The model adapts its conditioning strategy based on scene continuity:

$$S_t = \theta\left(M_t,\, C_{\text{story}},\, C^t_{\text{shot}},\, S^*_{t-1}\right) \tag{1}$$

where S^*_{t-1} denotes the conditional visual context. Specifically, we employ Text-Memory-and-Image-to-Video (TMI2V) for intra-scene shots by setting S^*_{t-1} as the last frame of S_{t-1} to maintain temporal continuity. For inter-scene transitions, we switch to Text-and-Memory-to-Video (TM2V) by setting $S^*_{t-1} = \emptyset$, relying solely on M_t to preserve subject identity across disjoint contexts.

The framework consists of three components: a subject-aware data curation pipeline that provides unambiguous subject supervision, a dual-query memory mechanism that retrieves long-context sub-

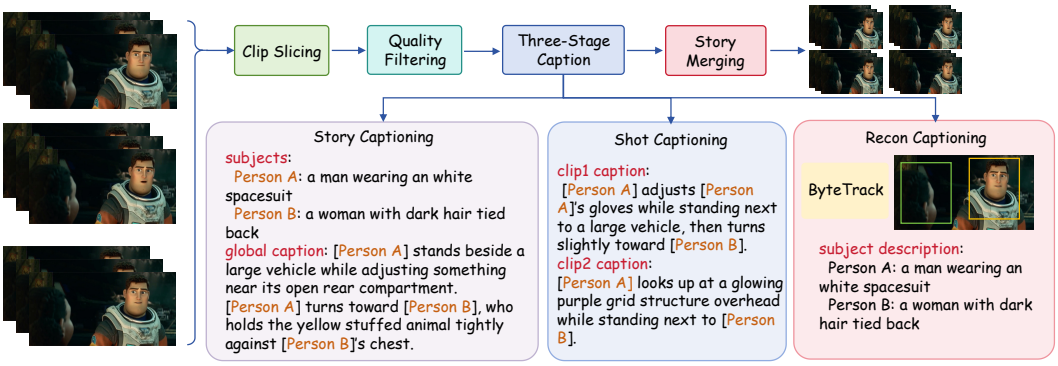

Figure 1: Overview of the subject-aware data curation pipeline. We design three-stage captions to support consistent long video generation. Story caption generates a global narrative and identifies all recurring subjects with fixed, pronoun-free names. Shot caption produces per-shot descriptions that adhere to the established subject inventory. Recon caption annotates selected subject-salient frames as reconstruction targets for memory-based identity supervision.

ject evidence and short-context generation cues, and a subject-anchored multi-task objective that jointly trains next-shot generation with memory-based subject reconstruction.

3.1 Subject-Aware Data Curation Pipeline

Memento requires training data that explicitly links *subject identity*, *cross-shot appearances*, and *reconstruction targets*. Unlike conventional video-caption datasets, which often describe subjects independently with ambiguous references such as the man," he," or "she," our curation maintains a consistent subject identity throughout the entire sequence. This enables reconstruction to directly supervise what subject information should be retained in memory.

As shown in Fig. 1, we first construct continuous multi-shot sequences from cinematic footage with standard quality filtering and subtitle removal. We then perform three-stage subject-aware annotation.

Stage 1: Story Captioning and Reconstruction Evidence Selection. We first process the full multi-shot sequence with Qwen3-VL Bai et al. (2025) to generate a global story caption and identify salient recurring subjects throughout the sequence. This stage establishes a subject inventory shared by all subsequent annotations, ensuring that each recurring subject is described with a fixed and explicit textual reference. In parallel, we apply ByteTrack Zhang et al. (2022) to track each subject over time. For each subject, we select the frame(s) where it occupies the largest visual area as reconstruction targets. These frames provide high-quality appearance evidence for the reconstruction branch.

Stage 2: Identity-Consistent Shot Captioning. Given the global story caption and the subject inventory from Stage 1, we caption each individual shot using Qwen3-VL. Instead of allowing generic noun phrases or pronouns, we enforce the fixed subject descriptions when referring to recurring subjects. This aligns local shot captions with the global subject identities and reduces ambiguity in cross-shot subject association, especially when multiple visually similar subjects appear in the same sequence.

Stage 3: Subject-Specific Reconstruction Captioning. Finally, we caption the selected reconstruction target frames from Stage 1 using Qwen3-VL. The generated captions are constrained to remain consistent with the global story caption and the established subject descriptions, so that each target image is paired with an unambiguous subject reference. These target image-caption pairs serve as ground truth for the memory-based subject reconstruction task. By optimizing this reconstruction objective, we force the model to actively retrieve and preserve identity-relevant cues from the memory bank, rather than relying solely on short-term visual context.

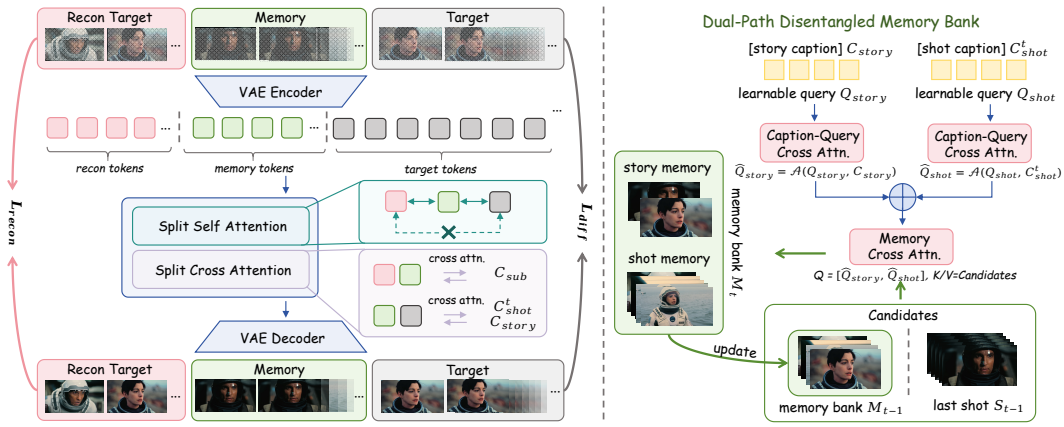

Figure 2: Overview of our framework. The left side illustrates the generation process, while the right side details the memory bank update mechanism. We employ split self-attention over overlapping local groups, allowing reconstruction-memory and memory-target interactions while avoiding full global attention. Split cross-attention injects both global story-level and local shot-level captions. Subject, story and shot captions condition separate learnable queries via caption-to-query cross-attention. The fused query retrieves relevant candidates from the memory bank and the last shot, and updates story and shot memory states for scalable long-form generation.

3.2 Dual-Path Disentangled Memory Bank

A fixed-size memory is necessary for scalable autoregressive long video generation, but a single memory selection criterion may entangle different types of historical information. In our setting, subject reconstruction requires long-range identity evidence, while next-shot generation relies more on recent shot context. As shown in the right part of Fig. 2, we therefore design a dual-path memory mechanism that retrieves subject-relevant and shot-relevant memories separately from a shared candidate pool.

Memory bank construction. At generation step t, Memento maintains a fixed-length memory bank M_t that summarizes useful historical visual information. Before generating the t-th shot, we form a candidate memory pool by combining the previous memory bank M_{t-1} with visual features extracted from the most recently generated shot S_{t-1}:

$$M_t^{\text{cand}} = [M_{t-1}\ ;\ \mathcal{E}(S_{t-1})] \in \mathbb{R}^{N_{\text{cand}} \times D} \tag{2}$$

where $\mathcal{E}(\cdot)$ denotes a pretrained VAE encoder that maps frames into latent features of dimension D, and N_{cand} is the number of candidate memory tokens. This construction allows the memory to retain selected information from earlier shots while continuously incorporating newly generated visual evidence.

Dual-query disentanglement. To separate identity-oriented retrieval from shot-specific retrieval, we introduce two sets of learnable query tokens: story queries $Q_{\text{story}} \in \mathbb{R}^{N_{\text{story}} \times D}$ and shot queries $Q_{\text{shot}} \in \mathbb{R}^{N_{\text{shot}} \times D}$. The story queries are conditioned on the global story caption to retrieve subject evidence that should remain stable across shots, while the shot queries are conditioned on the current shot caption to retrieve local references useful for the upcoming generation:

$$\hat{Q}_{\text{story}} = \mathcal{A}(Q_{\text{story}},\, \phi(C_{\text{story}}))\,, \qquad \hat{Q}_{\text{shot}} = \mathcal{A}\big(Q_{\text{shot}},\, \phi(C^t_{\text{shot}})\big) \tag{3}$$

where $\phi(\cdot)$ encodes text conditions into the shared latent space and $\mathcal{A}(\cdot,\cdot)$ denotes cross-attention. Through this conditioning, $\hat{Q}_{\text{story}}$ is guided by the global narrative and recurring subject descriptions, whereas $\hat{Q}_{\text{shot}}$ focuses on the local semantics of the current shot.

Adaptive memory selection. The two query sets independently score the candidate memory pool:

$$s_{\text{story}} = \mathcal{S}\Big(\hat{Q}_{\text{story}},\, M_t^{\text{cand}}\Big)\,, \qquad s_{\text{shot}} = \mathcal{S}\Big(\hat{Q}_{\text{shot}},\, M_t^{\text{cand}}\Big) \tag{4}$$

where $\mathcal{S}(\cdot,\cdot)$ denotes the relevance scoring function. We then select the top-K memory tokens for each path and concatenate the selected subsets to form the updated memory bank:

$$M_t = \big[\text{TopK}\big(M_t^{\text{cand}},\, s_{\text{story}}\big)\ ;\ \text{TopK}\big(M_t^{\text{cand}},\, s_{\text{shot}}\big)\big] \tag{5}$$

where $\mathrm{TopK}(M, s)$ retrieves the top-K tokens from M according to score s. The story-selected memories provide long-context subject evidence for identity grounding and reconstruction, while the shot-selected memories provide short-context visual cues for local generation. By decoupling the two retrieval paths, the memory bank reduces competition between persistent identity information and transient scene context.

3.3 Subject-Anchored Multi-Task Training

While the dual-path memory bank retrieves both long-range subject and short-term contextual cues, next-shot generation alone does not ensure the model actively utilizes subject cues, often favoring easier short-term context. As shown in the left part of Fig. 2, split self-attention and split cross-attention isolate memory-based subject reconstruction from next-shot generation, preventing the two tasks from interfering with each other while ensuring that reconstruction uses no additional visual cues. By forcing the model to recover subjects solely from memory, this objective compels it to learn the capability of retrieving and leveraging fine-grained subject features, rather than treating them as passive byproducts. We therefore introduce subject-anchored multi-task training, which couples standard next-shot video denoising with memory-based subject reconstruction. This auxiliary objective shares the memory-conditioned generator with next-shot denoising, adding identity supervision without extra training stages or inference-time overhead.

Next-shot generation. The primary task is conditional diffusion denoising for the current shot, using a flow-matching objective. Let $z_t = \mathcal{E}(S_t)$ denote the clean latent of the current shot. At noise level $\tau \in [0, 1]$, we construct the noisy shot latent $z_t^\tau = (1-\tau)z_t + \tau\epsilon_{\mathrm{vid}}$, where $\epsilon_{\mathrm{vid}} \sim \mathcal{N}(0, I)$. Given z_t^τ, the model predicts the target velocity $u_t = \epsilon_{\mathrm{vid}} - z_t$ conditioned on the memory bank M_t, the global story caption C_{story}, and the current shot caption C_{shot}^t:

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}\Big[\big\| u_t - v_\theta\big(z_t^\tau, \tau, M_t, C_{\mathrm{story}}, C_{\mathrm{shot}}^t\big) \big\|_2^2 \Big] \tag{6}$$

where v_θ denotes the velocity predictor, and the expectation is taken over training examples, sampled noise levels, and Gaussian noise. This objective trains the model to generate the next shot from both textual instructions and historical visual memory.

Subject-role reconstruction. To make identity preservation directly optimizable, we formulate an auxiliary Text-and-Memory-to-Image (TM2I) reconstruction task. Given a target subject image I_{sub}, we encode it as $z_{\mathrm{sub}} = \mathcal{E}(I_{\mathrm{sub}})$ and construct its noisy latent $z_{\mathrm{sub}}^\tau = (1-\tau)z_{\mathrm{sub}} + \tau\epsilon_{\mathrm{sub}}$, where $\epsilon_{\mathrm{sub}} \sim \mathcal{N}(0, I)$. The shared model predicts the target velocity $u_{\mathrm{sub}} = \epsilon_{\mathrm{sub}} - z_{\mathrm{sub}}$ conditioned only on the memory bank M_t and the subject description C_{sub}, without using any direct visual prompt:

$$\mathcal{L}_{\mathrm{recon}} = \mathbb{E}\Big[\big\| u_{\mathrm{sub}} - v_\theta\big(z_{\mathrm{sub}}^\tau, \tau, M_t, C_{\mathrm{sub}}\big) \big\|_2^2 \Big] \tag{7}$$

Since the reconstruction branch cannot access the target image except through its noisy latent, this objective encourages the model to retrieve identity-relevant appearance cues from memory under the guidance of the subject description. This discourages the memory from serving only short-term generation context and anchors it to recurring subject identities.

Total loss. The final training objective combines next-shot generation and subject reconstruction:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{diff}} + \lambda\, \mathcal{L}_{\mathrm{recon}} \tag{8}$$

where λ is a balancing hyperparameter that controls the trade-off between scene progression and subject reconstruction fidelity.

4 Experiments

We conduct extensive experiments to evaluate the effectiveness of Memento in generating long videos with high subject consistency and temporal coherence. First, we provide a comprehensive evaluation against state-of-the-art multi-shot video generation methods through quantitative metrics, qualitative visual results on the STbench and VBench Huang et al. (2024b), and a user study (Sec.4.1). Then, we investigate the individual contributions of our core components in Sec.4.2. Implementation details on model training and inference are provided in the appendix (Sec. A). We also show advanced ability of Memento in Sec.D.

Table 1: Quantitative comparison of video generation capabilities on STbench. We report aesthetic quality, semantic consistency at both story and shot levels, background consistency, and subject consistency across different granularities. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	Subject Consistency			Background Consistency	Semantic Consistency		Aesthetic
	Inter-shot	Intra-shot	Inter-scene		Story	Shot	
StoryDiffusion + Wan2.2-I2V Zhou et al. (2024)	0.5525	0.8448	0.6732	0.9767	0.2671	0.2689	**0.5310**
Mask²DiT Qi et al. (2025)	0.5994	0.8310	0.6204	0.9758	0.2536	0.2653	0.4820
StoryMem Zhang et al. (2025a)	0.6606	0.8146	0.6692	0.9732	0.2793	0.2681	0.4937
HoloCine Meng et al. (2025)	0.5791	0.8128	0.6594	0.9770	0.2720	0.2854	0.4568
Ours	**0.7338**	**0.8578**	**0.7268**	**0.9805**	**0.3063**	**0.2893**	0.4977

Table 2: Quantitative comparison of video generation capabilities on VBench. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	Aesthetic	Semantic Consistency		Background Consistency	Subject Consistency		
		Story	Shot		Inter-shot	Intra-shot	Inter-scene
Mask²DiT Qi et al. (2025)	0.5346	**0.3103**	**0.2931**	0.9803	0.9354	0.9503	0.9052
StoryMem Zhang et al. (2025a)	0.5366	0.2770	0.2467	0.9877	0.9361	0.9420	0.9544
HoloCine Meng et al. (2025)	0.5309	0.2817	0.2666	0.9801	0.7949	0.8655	0.8224
Ours	**0.5742**	0.2819	0.2677	**0.9902**	**0.9457**	**0.9552**	**0.9626**

4.1 COMPARISON

4.1.1 QUANTITATIVE RESULTS

We evaluate generated videos along four dimensions: subject consistency, background consistency, semantic consistency, and aesthetic quality. Subject consistency is measured at three temporal granularities: intra-shot, inter-shot, and inter-scene, capturing appearance stability within individual shots and across shots or scenes containing the same recurring subject. Detailed metric definitions are provided in Appendix F.

We compare Memento with representative methods from the three main paradigms: StoryDiffusion + Wan2.2-I2V Zhou et al. (2024); Wan et al. (2025) for storyboard-based generation, StoryMem Zhang et al. (2025a) for memory-conditioned generation, Mask²DiT Qi et al. (2025) and HoloCine Meng et al. (2025) for joint multi-shot generation. For HoloCine, we compute shot and scene level metrics at their corresponding granularities to avoid biases from temporal slicing.

As shown in Table 1, Memento ranks first on all three subject-consistency metrics, with the largest margin in inter-shot consistency (0.7338 versus 0.6606 for the strongest baseline). It also leads in story and shot level semantic consistency and background consistency, while ranking second in aesthetic quality, supporting stronger identity preservation with competitive generation quality.

To evaluate out-of-distribution (OOD) generalization beyond the cinematic training domain, we additionally test Memento on VBench Huang et al. (2024b). We use Qwen3-VL-8B to expand the original VBench prompts and provide the same expanded prompts to all compared methods, ensuring a consistent textual input for fair comparison. As shown in Table 2, Memento achieves the best results on all three subject-consistency metrics, background consistency, and aesthetic quality, while ranking second on both semantic-consistency metrics. These results support its ability to preserve subject identity and visual coherence in the OOD evaluation setting while maintaining competitive text alignment.

4.1.2 QUALITATIVE RESULTS

Fig. 3 compares long multi-shot generation results from StoryMem, HoloCine, and Memento, with red boxes highlighting cross-shot subject inconsistencies.

As shown in Fig. 3, HoloCine exhibits identity changes and character confusion across shots. StoryMem also shows appearance drift: the PhD student's hair changes from black to blond, while Robinson's appearance changes across shots, eventually depicting a different subject.

Memento preserves the PhD student's dark hair, glasses, and facial appearance, as well as Robinson's identity across scene transitions and interactions with Friday. The yellow-boxed reconstruc-

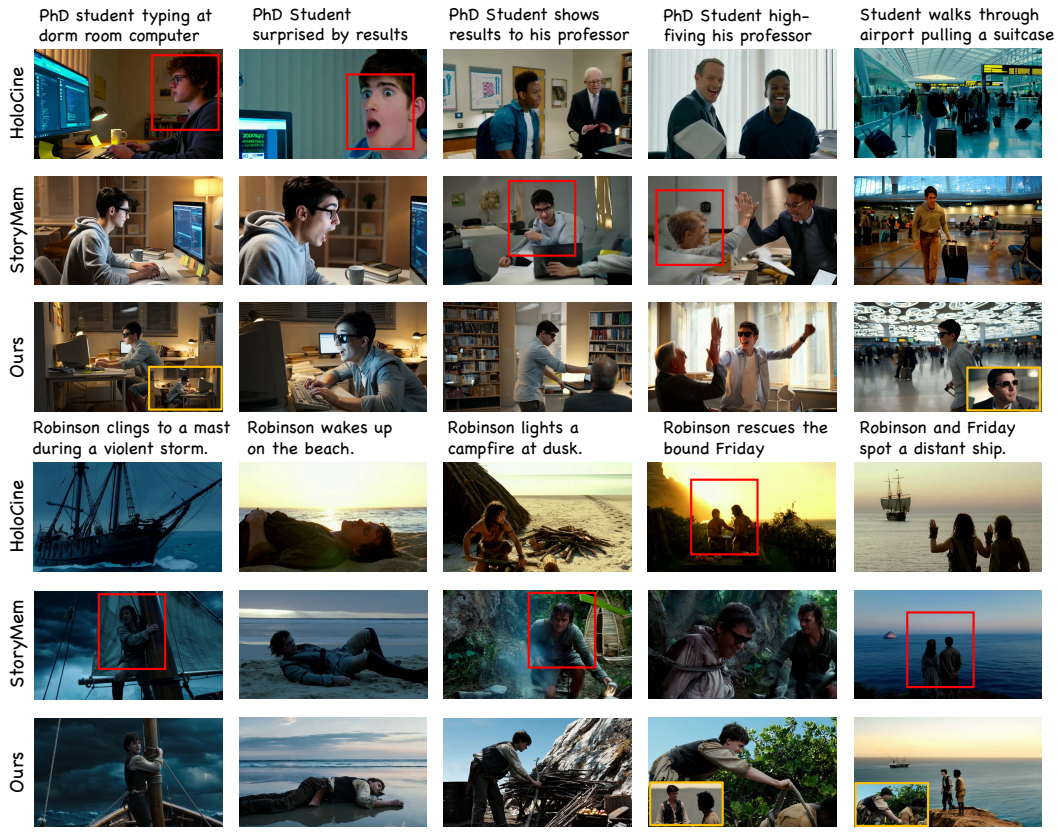

Figure 3: Qualitative comparison of multi-shot video generation against StoryMem Zhang et al. (2025a) and HoloCine Meng et al. (2025). Each row shows one method and each column shows a key story moment. Red boxes highlight cross-shot subject inconsistencies, while yellow boxes mark reconstruction frames. Our method better preserves subject identity across diverse scenes, viewpoints, and story progression.

tions match the generated subjects even in multi-character and long-shot settings, demonstrating subject reconstruction from memory. Additional qualitative comparisons on VBench are provided in Appendix C (Fig. 5).

4.1.3 User Study

To complement automatic evaluation, we conduct a user study with 10 participants on 30 diverse generation cases. For each case, evaluators perform pairwise comparisons between our method and each baseline, including StoryMem Zhang et al. (2025a) and HoloCine Meng et al. (2025). The comparisons are conducted under three criteria: cross-shot consistency, prompt following, and aesthetic quality. For each criterion and each baseline, evaluators choose whether our method is better, comparable, or worse, corresponding to Win, Tie, and Lose, respectively. The reported percentages are computed over 300 pairwise judgments, corresponding to 30 cases evaluated by 10 participants.

As shown in Fig. 4, our method is preferred over both baselines across all three criteria. Cross-shot consistency win rates reach 57.3% against StoryMem and 69.0% against HoloCine, and prompt following also shows clear advantages. Aesthetic quality receives favorable preferences alongside substantial tie rates. These results suggest improved perceived cross-shot consistency and prompt adherence with competitive or better visual quality.

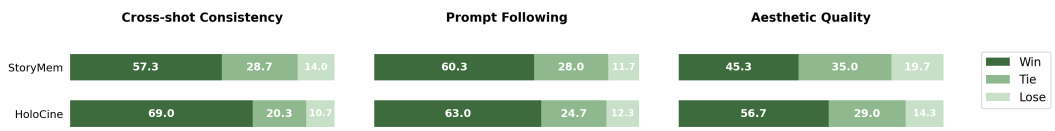

Figure 4: User study results. We report pairwise human preference comparisons between our method and each baseline over 30 generation cases evaluated by 10 participants. For each criterion, Win, Tie, and Lose indicate the percentage of comparisons where our method is preferred over, judged comparable to, or preferred less than the corresponding baseline.

Table 3: Ablation studies on the individual components of Memento. Starting from the baseline, we add each module and report aesthetic quality, semantic consistency (story and shot-level), background consistency, and subject consistency across different granularities. The best results are highlighted in **bold**.

Method	Subject Consistency			Background Consistency	Semantic Consistency		Aesthetic
	Inter-shot	Intra-shot	Inter-scene		Story	Shot	
Baseline	0.6606	0.8146	0.6692	0.9732	0.2793	<u>0.2681</u>	0.4937
Baseline + shot query	0.7227	0.7616	0.7177	<u>0.9752</u>	0.2793	0.2622	<u>0.4975</u>
Baseline + story query	0.7235	0.8235	0.7250	0.9633	0.2782	0.2620	0.4970
Baseline + shot query + recon. task	0.7319	0.7243	0.7082	0.9723	<u>0.2799</u>	0.2679	0.4901
Baseline + story query + recon. task	**0.7353**	**0.8649**	**0.7336**	0.9714	0.2571	0.2651	0.4974
Baseline + shot & story queries + recon. task	<u>0.7338</u>	<u>0.8578</u>	<u>0.7268</u>	**0.9805**	**0.3063**	**0.2893**	**0.4977**

4.2 Ablation Studies

We construct controlled variants from the same baseline by evaluating the shot and story query paths separately, pairing each path with the reconstruction objective, and finally combining all components under the same fixed memory budget. The results are reported in Tab. 3.

The single-query variants reveal the complementary roles of the two retrieval paths. The shot query provides modest gains in aesthetic quality and background consistency, consistent with its role in retrieving recent visual context. The story query yields larger gains in inter-shot and inter-scene subject consistency, supporting long-range identity preservation.

The reconstruction objective is most effective when coupled with story-conditioned retrieval. Adding reconstruction to the story-query variant improves all three subject-consistency metrics, with the largest gain of 0.0414 occurring in intra-shot consistency. Its effect is not uniform across retrieval configurations: with the shot-query variant, reconstruction improves inter-shot consistency by 0.0092 but does not improve the other two subject metrics. These results suggest that reconstruction is particularly effective when the retrieved memory is already oriented toward long-range subject evidence.

Finally, the full model provides the best overall balance rather than maximizing every individual subject metric. Combining both query paths with reconstruction achieves the strongest aesthetic quality, story- and shot-level semantic consistency, and background consistency among the ablations, while ranking second on all three subject-consistency metrics. This result supports the intended division of labor: story-conditioned retrieval anchors recurring identities, whereas shot-conditioned retrieval preserves the contextual evidence needed for coherent next-shot generation.

5 Conclusion and Discussion

We present Memento, a subject-reconstruction-guided framework for consistent long video generation. By combining dual-query memory retrieval with memory-based subject reconstruction, Memento explicitly preserves long-term identity evidence while maintaining short-term contextual continuity for next-shot generation. Together with a subject-aware data curation pipeline, Memento achieves strong cross-shot subject consistency, narrative coherence, and visual quality, while naturally extending to long-form generation without architectural modification. Memento has two limitations. First, autoregressive generation may propagate errors when degraded shots are written into memory. Second, physical plausibility remains limited by the video backbone, which lacks explicit modeling of gravity, object permanence, and rigid-body dynamics. Future work will study robust memory filtering and physics-aware generation.

AI DISCLOSURE

We used Qwen3-VL-8B to generate captions as part of the subject-aware data curation pipeline described in this paper. Separately, we used generative AI tools to assist with language editing and polishing of the manuscript. All AI-assisted manuscript text was carefully reviewed and revised by the authors, who take full responsibility for the accuracy, originality, and scientific integrity of the final paper.

ETHICS STATEMENT

This work uses cinematic footage to train a generative video model and includes a user study with 10 participants. We report the user study results in aggregate. The use of source footage and the generation of recognizable subjects raise considerations concerning intellectual property, likeness rights, and deceptive synthetic media. Applications of this method should respect the rights associated with source material and clearly identify generated content where appropriate.

REPRODUCIBILITY STATEMENT

We describe the data curation pipeline, model architecture, and training objectives in the main paper. The appendix provides dataset statistics, memory selection details, training and inference settings, and definitions of the evaluation metrics. We also report the user study design. These details support replication of the experiments, although reproducing the exact results may depend on access to the source footage, pretrained models, and the reported computing resources.

REFERENCES

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. *arXiv preprint arXiv:2512.07802*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

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.

Google DeepMind. Veo 3 video model. `https://deepmind.google/models/veo/`, 2026.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17281–17291, 2025.

Jingwen He, Hongbo Liu, Jiajun Li, Ziqi Huang, Qiao Yu, Wanli Ouyang, and Ziwei Liu. Cut2next: Generating next shot via in-context tuning. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 2024.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

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.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28405–28415, 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 preprint arXiv:2412.03603*, 2024.

Kuaishou. Kling video model. `https://kling.kuaishou.com`, 2026.

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.

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

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.

OpenAI. Sora 2 video model. `https://openai.com/research/sora-2`, 2026.

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.

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

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Maskˆ 2dit: Dual mask-based diffusion transformer for multi-scene long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18837–18846, 2025.

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.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

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.

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. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. *arXiv preprint arXiv:2508.11484*, 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. In *International Conference on Learning Representations*, volume 2026, pp. 83349–83375, 2026.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788*, 2024.

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

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.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025a.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Yuang Zhang, Junqi Cheng, Haoyu Zhao, Jiaxi Gu, Fangyuan Zou, Zenghui Lu, and Peng Shu. Shouldershot: Generating over-the-shoulder dialogue videos. *arXiv preprint arXiv:2508.07597*, 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 sequence. *arXiv preprint arXiv:2407.16655*, 2024.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2412.02259*, 3(6), 2024.

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.

A Implementation Details

Training Setup. Our video generation backbone is built upon Wan2.2 Wan et al. (2025) 14B. The training data is processed through our subject-aware data curation pipeline, resulting in 20K clips at a resolution of 480×832. We employ the AdamW Loshchilov & Hutter (2017) optimizer with a base learning rate of $1e-5$. The model is trained on 16 NVIDIA H100 GPUs. The loss weight parameter λ in Eq. 8 is empirically set to 0.5. To facilitate robust classifier-free guidance (CFG) during inference, we randomly drop the textual conditions C_{story}, C_{shot} and C_{sub} with a probability of 10% during training.

Inference Configurations. We follow the STbench autoregressive evaluation protocol and adopt the scene transition strategy of StoryMem Zhang et al. (2025a). The first shot S_1 is generated from Gaussian noise with standard T2V. For later shots, we use TMI2V within the same scene, injecting the last frame of the previous shot for pixel-level continuity, and switch to TM2V across scene changes, relying on text prompts and the updated memory bank M_t for long-term identity preservation. We use the UniPC Zhao et al. (2023) sampler with 40 denoising steps for both generation and reconstruction, and set the CFG scale for global and local text conditions to 3.5.

B Dual-Query Keyframe Selection

We detail the scoring function $\mathcal{S}$ and the top-K operator in Sec. 3.2. We retain the terms *shot* and *story* for the local/context and global/identity retrieval paths, respectively. Selection operates on complete memory frames: spatial tokens are used to compute relevance, but are not selected independently. The number of learnable queries and the number of retained frames are distinct: we use $N_{\text{shot}} = 8$ and $N_{\text{story}} = 2$ queries, with frame budgets $K_{\text{shot}} = 6$ and $K_{\text{story}} = 2$.

Candidate memory pool. For shot t, write the candidate pool as $M_t^{\text{cand}} = \{(m_f, a_f)\}_{f=1}^{F}$, where $m_f \in \mathbb{R}^{16\times H_z \times W_z}$ is a VAE-encoded memory frame and a_f is its absolute frame index on the 16-fps video timeline. During training, historical shots $S_1, \ldots, S_{t-1}$ are sampled at 4 fps, retaining at most eight frames per shot; these historical candidates are then randomly subsampled to at most eight frames in total. We concatenate them with eight temporally uniform samples from S_{t-1}, giving at most 16 candidate entries. The two sources may contain samples from the same historical shot. During inference, the memory persists across shots. After completing a shot, eight uniformly sampled frames are VAE-encoded and appended to the retained bank. Selection is triggered only when the pool exceeds the eight-frame capacity; otherwise, all available entries are retained.

Text-conditioned queries and visual scoring. A dedicated 3D convolution with patch size $(1, 2, 2)$ projects each 16-channel memory latent into P spatial tokens of dimension D. We denote these tokens by $X_{f,p} \in \mathbb{R}^D$, where $p \in \{1, \ldots, P\}$. Thus, D denotes the query/DiT feature width, rather than the number of VAE latent channels. To specify the text inputs to the two paths, let C_{id} denote the global subject description used for identity retrieval. The implemented caption-to-query attention is

$$\begin{aligned} \hat{Q}_{\text{shot}} &= \mathcal{A}\big(Q_{\text{shot}}, \phi([C_{\text{story}}; C^t_{\text{shot}}])\big)\,, \\ \hat{Q}_{\text{story}} &= \mathcal{A}(Q_{\text{story}}, \phi(C_{\text{id}}))\,, \end{aligned} \tag{9}$$

where ϕ is the T5 text encoder. The shot path receives the full global-and-shot caption, whereas the story path receives only the global identity description. Both paths search the same candidate pool; their short- and long-context roles arise from their conditioning, rather than separate temporal search windows.

The text-conditioned queries next cross-attend to the memory tokens with rotary positional encoding (RoPE), producing visually enhanced queries $\tilde{Q}_{\text{shot}}$ and $\tilde{Q}_{\text{story}}$. A further query–key attention computation scores the memory tokens using the enhanced queries and RoPE-encoded memory keys. Let $A^h_{q,f,p}$ denote the resulting attention score for query q, frame f, spatial position p, and head h. Averaging over the N_h heads and spatial positions gives

$$R_{q,f} = \frac{1}{N_h P} \sum_{h=1}^{N_h} \sum_{p=1}^{P} A^h_{q,f,p}, \qquad R \in \mathbb{R}^{(N_{\text{shot}}+N_{\text{story}})\times F}. \tag{10}$$

We use R to avoid overloading S_t, which denotes a video shot. The two frame-level relevance scores in Sec. 3.2 are

$$s_b(f) = \frac{1}{N_b} \sum_{q \in \mathcal{Q}_b} R_{q,f}, \qquad b \in \{\text{shot}, \text{story}\}, \tag{11}$$

where $\mathcal{Q}_b$ indexes the queries of path b.

Budgeted selection without overlap. The shot path selects its six highest-scoring candidates first. We then exclude these candidate indices from the story path and select two additional frames:

$$\begin{aligned} \mathcal{I}_{\text{shot}} &= \text{argTopK}_{K_{\text{shot}}}(s_{\text{shot}}), \\ \bar{s}_{\text{story}}(f) &= \begin{cases} -\infty, & f \in \mathcal{I}_{\text{shot}}, \\ s_{\text{story}}(f), & f \notin \mathcal{I}_{\text{shot}}, \end{cases} \\ \mathcal{I}_{\text{story}} &= \text{argTopK}_{K_{\text{story}}}(\bar{s}_{\text{story}}). \end{aligned} \tag{12}$$

Here argTopK_k returns the indices of the k largest entries. Consequently, $\mathcal{I}_{\text{shot}} \cap \mathcal{I}_{\text{story}} = \emptyset$, and the updated memory is

$$M_t = \left[(m_f)_{f \in \mathcal{I}_{\text{shot}}} \; ; \; (m_f)_{f \in \mathcal{I}_{\text{story}}} \right]. \tag{13}$$

The no-overlap constraint concerns candidate entries; it does not impose visual dissimilarity between frames. When fewer than eight entries are available, the selector retains the available entries without requiring a full eight-frame bank.

Straight-through training. We use a straight-through surrogate to train the discrete selector through the downstream flow-matching objective. Let i_j be the hard-selected candidate index for output slot j, and let $b(j)$ identify its retrieval path. We compute

$$p_{\text{shot}} = \text{softmax}(s_{\text{shot}}/T_{\text{sel}}), \qquad p_{\text{story}} = \text{softmax}(\bar{s}_{\text{story}}/T_{\text{sel}}), \tag{14}$$

with $T_{\text{sel}} = 0.1$, distinct from the diffusion noise level τ in Sec. 3.3. With sg denoting stop-gradient, the selected-entry weighting is

$$\begin{aligned} w_{j,f} &= \mathbf{1}[f = i_j] \left(1 - \text{sg}\big(p_{b(j)}(f)\big) + p_{b(j)}(f)\right), \\ \tilde{m}_j &= \sum_{f=1}^{F} w_{j,f} m_f. \end{aligned} \tag{15}$$

In the forward pass, $w_{j,f} = \mathbf{1}[f = i_j]$ and $\tilde{m}_j = m_{i_j}$, so the generator receives exactly the hard-selected frames. In the backward pass, gradients reach the scoring network through the selected softmax probabilities. This is a surrogate gradient; the discrete top-K indices themselves are not differentiated. Although only selected entries receive nonzero weights, their softmax probabilities depend on all unmasked candidate scores. The story softmax uses the local-selection mask in Eq. 12, so its scoring gradient does not reach entries already assigned to the shot path. The learnable queries, patch embedding, and cross-attention modules are optimized end-to-end through the downstream objective in Eq. 8, without explicit keyframe-selection labels.

Inference and temporal indexing. At evaluation time, the selector returns strictly one-hot selection rows. The inference pipeline extracts their argmax indices and removes duplicates. It optionally reserves an initial prefix of memory frames as temporal anchors; the default anchor count is zero. If fewer than six local entries remain after deduplication, missing entries are filled according to the maximum query score for each candidate. The global entries are appended after the local entries in ascending index order, maintaining the disjoint selection. In distributed inference, the selected indices are broadcast across ranks to prevent small floating-point differences from producing inconsistent memory banks.

Each selected frame retains its absolute timestamp a_f, independent of its position in the concatenated memory. For RoPE lookup, timestamps are converted to relative latent-frame offsets with respect to the current shot, clipped to $[-512, 511]$, and shifted by the lookup-table center 512. The same temporal-index convention is used during training and inference.

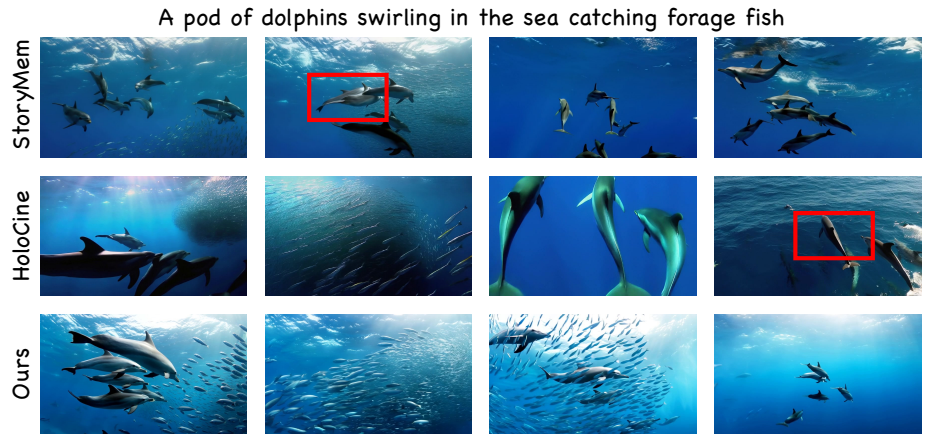

Figure 5: Qualitative comparison of video generation against StoryMem Zhang et al. (2025a) and HoloCine Meng et al. (2025) on VBench. Each row shows one method and each column shows a key moment.

Conditioning the shared DiT. The selected clean memory latents are concatenated along the temporal dimension with the noisy target latents. In joint training, the sequence is arranged as $[z^{\tau}_{\text{sub}}; M_t; z^{\tau}_t]$, with the subject-reconstruction and video-generation interactions governed by the split-attention scheme in Sec. 3.3. Memory positions act only as clean conditions: they are not noised, their denoising/loss-mask entries are zero, and they are excluded from the reconstruction targets. RoPE positions follow the stored timestamps rather than the selection-slot order. The noisy target positions are trained with the flow-matching losses defined in the main text.

C Additional Experiment

To complement the quantitative results reported in the main text, we provide qualitative comparisons on VBench Huang et al. (2024b) in Figure 5, highlighting Memento's ability to maintain multi-subject coherence in complex dynamic scenes. These comparisons reveal that baseline methods struggle to preserve distinct and consistent subjects. StoryMem fails to distinguish individual instances, leading to artifact-prone subject merging. HoloCine generates physically inconsistent frames with distorted body structures. Conversely, Memento demonstrates robust stability, accurately maintaining the number of subjects (cardinality) and preserving distinct, realistic identities even in complex dynamic scenes. This confirms the generality of our method in handling diverse video generation benchmarks beyond long-form narratives.

D Advanced Capabilities

Beyond standard multi-shot generation, Memento exhibits two advanced capabilities, as shown in Fig. 6. First, it supports age-consistent identity preservation: the model can generate narratives where the same character changes age while maintaining stable identity cues such as facial structure, hairstyle tendency, and overall appearance. This suggests that Memento does not merely copy a fixed reference, but separates persistent identity from controllable temporal variations. Second, our autoregressive generation with dynamic memory update scales naturally to long-form synthesis. In a 5-minute generation example, Memento maintains coherent subject appearance, scene progression, and visual quality across many connected shots, demonstrating practical scalability toward minute-level narrative video generation.

E Additional Data Details

We provide additional statistics of the curated training data in this section. The dataset contains 2,033 video sequences and 20,227 clips in total. Each sequence contains 9.95 clips on average, with



Figure 6: Advanced capabilities of Memento. Top: our method can generate narratives involving age variation while preserving a recognizable subject identity across different life stages. Bottom: our autoregressive multi-shot generation framework scales to long-form video synthesis; we further generate a 5-minute video to demonstrate its ability to maintain coherent story progression and subject consistency over extended duration.

a minimum of 6 clips and a maximum of 16 clips. The average sequence duration is 45.9 seconds, the median duration is 42.9 seconds, and the duration ranges from 30 to 65 seconds.

For data annotation, we use Qwen3-VL-8B for caption generation, PaddleOCR for OCR detection, and ByteTrack for identifying reconstruction frames. Qwen3-VL-8B is called 26,252 times in total across three captioning tasks: sequence captioning, clip captioning, and reconstruction captioning. The detailed caption statistics are shown in Table 4.

Table 4: Statistics of the curated dataset and generated captions. Caption length is measured by character count.

Type	Count	Min	Max	Mean	Median
Story caption	2,030	80	1,485	543	513
Shot caption	20,189	152	999	459	450
Subject description	3,992	33	641	130	117

F Evaluation Metrics

We report seven metrics in four categories: subject consistency (inter-shot, intra-shot, and inter-scene), background consistency, semantic consistency (scene and shot levels), and aesthetic quality. Higher values indicate better performance for all seven metrics. A shot is a continuous segment, a scene groups shots belonging to the same scene, and a story comprises the full multi-scene sequence. We distinguish these evaluation units from the 8-frame clips used as inputs to ViCLIP. Unless otherwise specified, feature vectors are L2-normalized before cosine similarity is computed.

F.1 Subject Consistency

Subject consistency (SC) measures visual similarity within shots and across shots or scenes involving recurring characters. We use the subscripts in, sh, and sc for intra-shot, inter-shot, and inter-scene scores, respectively. The three metrics below use frame- or video-level representations as proxies for subject consistency; these representations may also reflect background, pose, and action. Character grouping restricts which shots or scenes are compared, but does not by itself isolate a character's visual features.

F.1.1 INTRA-SHOT SUBJECT CONSISTENCY

For a shot s with $T_s \geq 2$ evaluated frames $f_1, \ldots, f_{T_s}$, we extract DINOv2 Oquab et al. (2023) features and denote their L2-normalized representations by $\hat{\mathbf{d}}_j$. The intra-shot subject consistency score combines similarity between adjacent evaluated frames and similarity to the first frame:

$$\mathrm{SC}_{\mathrm{in}}(s) = \frac{1}{2}\left(\frac{1}{T_s - 1}\sum_{j=1}^{T_s-1} \hat{\mathbf{d}}_j^\top \hat{\mathbf{d}}_{j+1} + \frac{1}{T_s}\sum_{j=1}^{T_s} \hat{\mathbf{d}}_1^\top \hat{\mathbf{d}}_j\right). \tag{16}$$

The two terms receive equal weight. This metric measures appearance stability within a shot, rather than consistency across shot boundaries.

F.1.2 INTER-SHOT SUBJECT CONSISTENCY

To evaluate consistency across shots, we group shots by character identity. Let $\mathcal{G}_c = \{s_1, \ldots, s_{N_c}\}$ denote the shots containing character c. Character identities are understood within their corresponding stories.

For each shot s_i, we average its ViCLIP Wang et al. (2024) clip features and L2-normalize the result to obtain $\hat{\mathbf{v}}_i$. The inter-shot subject consistency for character c is the average cosine similarity over all distinct shot pairs in $\mathcal{G}_c$, including nonconsecutive shots:

$$\mathrm{SC}_{\mathrm{sh}}(c) = \frac{2}{N_c(N_c - 1)} \sum_{1 \leq i < j \leq N_c} \hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j. \tag{17}$$

The reported *Inter-shot* score is the mean of $\mathrm{SC}_{\mathrm{sh}}(c)$ over character groups with $N_c \geq 2$; groups with fewer than two shots are excluded.

F.1.3 INTER-SCENE SUBJECT CONSISTENCY

This metric evaluates similarity across different scenes containing the same character. Let $\mathcal{H}_c = \{r_1, \ldots, r_{M_c}\}$ denote the scenes containing character c.

Let $\hat{\mathbf{q}}_k$ be the L2-normalized ViCLIP representation of scene r_k, obtained by averaging its clip features. The inter-scene subject consistency for character c is the average cosine similarity over all distinct scene pairs:

$$\mathrm{SC}_{\mathrm{sc}}(c) = \frac{2}{M_c(M_c - 1)} \sum_{1 \leq i < j \leq M_c} \hat{\mathbf{q}}_i^\top \hat{\mathbf{q}}_j. \tag{18}$$

The reported *Inter-scene* score is the mean of $\mathrm{SC}_{\mathrm{sc}}(c)$ over character groups with $M_c \geq 2$; groups appearing in only one scene are excluded.

F.2 BACKGROUND CONSISTENCY

Background consistency measures visual stability within shots. For shot s, let $\hat{\mathbf{b}}_j$ denote the L2-normalized OpenAI CLIP ViT-L/14 feature used for background evaluation at frame j. We compute the average similarity between adjacent evaluated frames:

$$\mathrm{BC}(s) = \frac{1}{T_s - 1}\sum_{j=1}^{T_s-1} \hat{\mathbf{b}}_j^\top \hat{\mathbf{b}}_{j+1}. \tag{19}$$

For a scene r containing a set of shots $\mathcal{S}_r$, its background consistency score is $\mathrm{BC}(r) = |\mathcal{S}_r|^{-1} \sum_{s \in \mathcal{S}_r} \mathrm{BC}(s)$. Similarities are computed within individual shots; frames on opposite sides of a shot or scene boundary are not compared by this metric.

F.3 SEMANTIC CONSISTENCY

Semantic consistency is measured by text alignment (TA), with subscripts sc and sh denoting scene and shot levels. We use the OpenGVLab/ViCLIP-B-16-hf checkpoint of ViCLIP **?**. For an evaluated video unit, we average its 8-frame clip features and L2-normalize the result. Its alignment score is the cosine similarity between this video representation and the L2-normalized representation of its associated text.

Global semantic consistency (Story). For each generated scene r, we uniformly sample 24 frames by default and divide them into three 8-frame clips. We concatenate the shot prompts within that scene to form the associated text description. Let $\hat{\mathbf{q}}_r$ and $\hat{\mathbf{h}}_r$ denote the resulting normalized video and text representations. The scene-level alignment score, reported under the *Story* column in the tables, is

$$\mathrm{TA}_{\mathrm{sc}}(r) = \hat{\mathbf{q}}_r^\top \hat{\mathbf{h}}_r. \tag{20}$$

This metric evaluates alignment with the combined shot prompts within a scene; it does not encode the entire multi-scene story at once.

Shot-level semantic consistency. We evaluate each shot s independently by uniformly sampling frames and dividing them into 8-frame clips. Let $\hat{\mathbf{v}}_s$ denote the normalized shot-level video representation and $\hat{\mathbf{h}}_s$ the normalized text representation of that shot's prompt. The shot-level alignment score is

$$\mathrm{TA}_{\mathrm{sh}}(s) = \hat{\mathbf{v}}_s^\top \hat{\mathbf{h}}_s. \tag{21}$$

The reported *Shot* score is the mean of $\mathrm{TA}_{\mathrm{sh}}(s)$ over all evaluated shots.

F.4 Aesthetic Quality

Aesthetic quality measures the visual appeal of individual frames. For each frame f, we extract an OpenAI CLIP ViT-L/14 image feature, L2-normalize it, and feed it into the LAION aesthetic MLP predictor to obtain a raw score $a_{\mathrm{raw}}(f)$. We apply the following rescaling:

$$a(f) = \frac{a_{\mathrm{raw}}(f) - 1}{9}. \tag{22}$$

The reported *Aesthetic* score is the mean of $a(f)$ over all evaluated frames. This frame-level metric measures visual appeal and does not directly assess motion quality or cross-shot consistency.