FRAMEMORROW: FUTURE-GUIDED FRAME SELECTION WITH PROSPECTIVE TOKENS FOR LONG-HORIZON VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon video generation requires models to effectively leverage an increasingly long generation history. As the generated history grows, retaining all previous content becomes increasingly expensive and redundant, making effective historical selection essential. Existing approaches often determine historical relevance from the current content. However, information that is relevant to the present is not necessarily useful for future generation, while seemingly less relevant history may become important later. Our key insight is that *historical information should be selected according to its relevance to future information needs*. Capturing these needs does not require generating the full future. A compact representation of what becomes important next is sufficient to guide historical selection. Building on this insight, we propose **FRAMEMORROW**, a prospective frame selector that predicts a small set of *prospective tokens* representing future information needs and uses them to identify relevant information from history. FRAMEMORROW selects explicit historical frames rather than model-specific internal states, enabling plug-and-play integration across diverse generators, including closed-source models, with little additional inference cost. We evaluate FRAMEMORROW across five benchmarks and 11 generative models spanning long-video generation, interactive generation, and action-conditioned world models. Extensive experiments demonstrate improvements in long-range consistency, visual quality, and action alignment across diverse generation settings.

1 INTRODUCTION

Long-horizon video generation requires models not only to continuously extend visual content, but also to preserve coherent subjects, objects, and scenes over time (Yang et al., 2025; Zhang et al., 2025a). As generation proceeds, visual details generated earlier in the video gradually fall outside the model's limited input window (Hu et al., 2026). When such content becomes relevant again, subjects can change appearance, object states can become inconsistent, and revisited scenes can no longer match their earlier observations (Xiao et al., 2026). Keeping all historical information is impractical, as it increases computation and introduces a large amount of redundant context (Yi et al., 2025; Ji et al., 2025). Moreover, not all historical information is equally useful at every generation step (Ji et al., 2025; An et al., 2026). The key challenge is therefore to selectively preserve the historical information that matters for future generation.

Many existing methods identify useful history based on its relevance to the current content (Zhang et al., 2025b), often using the recent visual context to retrieve related information from the past (Hu et al., 2026; Ye et al., 2026; Wang et al., 2026a; Ding et al., 2026). Such a strategy mainly answers which historical information is most relevant to what is visible now. However, this can differ from what will actually be useful for future generation. Information that closely matches the present may provide little additional value for what comes next, while earlier information that appears less relevant now can become important again as the generation evolves. As illustrated in Fig. 1, matching the current view can favor visually similar history, while information that is less relevant to the present may better support the upcoming generation. Therefore, the importance of historical information should not be determined only by its relevance to the present, but also by its potential

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 1: **Motivation for future-relevant history selection.** Given the same available history and known next condition, current-context matching favors observations similar to the current view, whereas FRAMEMORROW selects earlier history that is more relevant to the upcoming generation. The selected history helps preserve previously observed visual details in the generated continuation.

usefulness for the future. *Can historical information be selected according to its relevance to the future rather than to the current content?*

However, the challenge is that the future has not been generated when the selection is made. Since the goal is to determine which historical information will be useful for future generation, there is no need to generate the full future itself. Instead, it is sufficient to predict a compact representation that captures future information needs and serves as a proxy for the future. Such a representation can then guide historical selection toward information that is likely to matter next.

Motivated by this, we propose FRAMEMORROW, which predicts a small set of *prospective tokens* as compact representations of future information needs. These tokens capture what potentially become important next and are used to identify historical information that is relevant to future generation. FRAMEMORROW then selects explicit historical frames based on this relevance. Because it selects explicit historical frames rather than model-specific internal states, the same selector can be used across different generators, while each model processes the selected frames in its native way. This makes FRAMEMORROW plug-and-play across different generators, including closed-source models. Moreover, predicting only a few prospective tokens and selecting only a few historical frames keeps FRAMEMORROW lightweight with little additional inference cost.

Our contributions are as follows:

- We formulate using future information needs as the condition for frame selection. Selection based only on current content cannot predict future information needs and retrieve frames that are relevant to the present but unhelpful for future generation.
- We propose FRAMEMORROW, which introduces *prospective tokens* to explicitly represent future information needs and identify relevant information from the history.
- We design FRAMEMORROW as a *plug-and-play*, lightweight selector that outputs explicit historical frames, enabling broad compatibility across different generators, including closed-source models, with little inference overhead.
- We extensively evaluate FRAMEMORROW across five benchmarks and 11 generative models, covering long-video generation, interactive generation, and action-conditioned world models, with improvements in long-range consistency, visual quality, and action alignment.

2 RELATED WORK

Long-Horizon Video Generation. Autoregressive video models extend videos by conditioning new content on previously generated outputs (Yin et al., 2025; Huang et al., 2026). CausVid (Yin et al., 2025) distills a bidirectional diffusion model into a few-step causal generator for streaming synthesis. Self-Forcing (Huang et al., 2026) reduces the training–inference gap by using self-generated

context during training. LongLive (Yang et al., 2025) supports real-time long-video generation with changing text prompts, while Matrix-Game 3.0 (Wang et al., 2026b) extends interactive world generation with action control and long-horizon memory. Rather than developing another generation backbone, our work provides a plug-and-play frame selector that supports diverse existing generators by supplying relevant historical information.

Historical Information Selection. Historical information can be retrieved as visual context or retained within a generator's internal cache. LongLive-RAG retrieves historical latents using the latest generated content (Hu et al., 2026), while DySink selects visually relevant historical frames as dynamic frame sinks (Ye et al., 2026). For narrative generation, MemFlow retrieves history using the upcoming chunk prompt (Ji et al., 2025), and Memento separately retrieves identity evidence and short-range shot cues (Wei et al., 2026). Mem-World combines planned actions with scene geometry to retrieve relevant historical observations (Zheng et al., 2026). For cache management, PaFu-KV learns token salience from a bidirectional teacher (Chen et al., 2026a), while Future Forcing constructs future-query proxies to guide cache eviction and merging (Luo et al., 2026a). Unlike existing approaches that determine historical relevance from observed content or model-specific memory states, our method selects history according to its relevance to future information needs.

Learning to Select Frames. Learned frame selection reduces redundant visual input for video understanding. Frame-Voyager learns query-conditioned frame combinations from rankings provided by a video-language model (Yu et al., 2025). FrameOracle predicts both query-relevant frames and an adaptive frame budget (Li et al., 2025). Related approaches learn selection through multimodal-model supervision (Hu et al., 2025), flexible selection policies (Buch et al., 2025), or reinforcement-learning rewards (Qin et al., 2026). These methods select evidence to answer questions or reason about an available video. Unlike frame selectors designed for fully observed videos, FRAMEMORROW predicts *prospective tokens* to represent information needs for content that has not yet been generated, enabling lightweight selection of relevant historical frames.

3 METHOD

Overview. FRAMEMORROW predicts a small set of *prospective tokens* to represent future information needs by given eligible history $\mathcal{H}_t$, recent context $\mathcal{L}_t$, and the known rollout condition c_t^+. These tokens score historical relevance and guide frame selection. During training, a frozen visual teacher ranks historical frames by their correspondence with future content, providing ranking supervision for the tokens. At inference, FRAMEMORROW outputs explicit historical frames, which augment a compatible frozen generator through its own conditioning mechanism (Fig. 2).

3.1 PROSPECTIVE FRAME SELECTION

At rollout step t, we distinguish the backbone's recent context from the eligible long-term history:

$$\mathcal{L}_t = \{x_{t-L+1}, \ldots, x_t\}, \qquad \mathcal{H}_t = \{x_{\tau_i}\}_{i=1}^{N}, \quad \tau_i \le t - L. \tag{1}$$

Both sets are observed by the selector, but only $\mathcal{H}_t$ is eligible for retrieval. Recent context thus informs which past evidence to recall without occupying the long-term memory budget. The condition c_t^+ is a text prompt, action sequence, or control signal available at step t. It contains no later user input, environment feedback, or future observation. For a memory budget K, we select

$$\hat{\mathbf{r}}_t = S_\theta(\mathcal{H}_t, \mathcal{L}_t, c_t^+) \in \mathbb{R}^N, \qquad \mathcal{I}_t = \mathrm{TopK}(\hat{\mathbf{r}}_t, K), \qquad \mathcal{M}_t = \mathcal{H}_t[\mathcal{I}_t]. \tag{2}$$

The selected frames are ordered by their timestamps before being passed to the backbone.

3.2 PROSPECTIVE TOKENS

To represent future information needs without generating future content, a compact causal Transformer predicts prospective tokens from the observed history, recent context, and rollout condition. A frozen visual encoder E_v and learned projection P_v encode each observed frame as $\mathbf{z}(x) = P_v[E_v(x)] \in \mathbb{R}^d$. We form the temporally ordered sequences $\mathbf{H}_t = [\mathbf{h}_1, \ldots, \mathbf{h}_N]$, with $\mathbf{h}_i = \mathbf{z}(x_{\tau_i})$, and $\mathbf{L}_t = [\mathbf{z}(x_{t-L+1}), \ldots, \mathbf{z}(x_t)]$. A frozen modality-specific encoder E_c and learned projection P_c produce $\mathbf{C}_t^+ = P_c[E_c(c_t^+)]$. Selector weights are shared across compatible backbones within each condition modality. Different modalities share the architecture and interface.

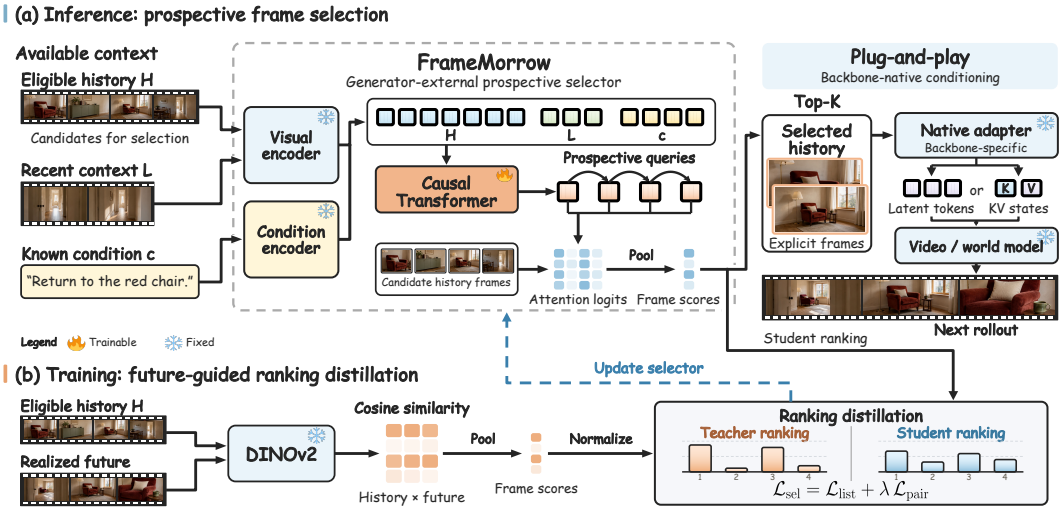

Figure 2: **Overview of FRAMEMORROW.** (a) At inference, the selector autoregressively predicts prospective tokens from history, recent context, and the rollout condition. These tokens score eligible historical frames and guide top-K selection. Each frozen generator processes the selected frames through its own conditioning mechanism. (b) During training, frozen DINOv2 features provide future-grounded frame rankings, which supervise the prospective tokens through listwise and pairwise ranking losses.

After processing $[\mathbf{H}_t, \mathbf{L}_t, \mathbf{C}_t^+]$, the selector autoregressively predicts M prospective tokens:

$$\mathbf{q}_t^m = F_\psi(\mathbf{H}_t, \mathbf{L}_t, \mathbf{C}_t^+, \mathbf{q}_t^{<m}), \qquad m = 1, \ldots, M. \tag{3}$$

All predicted tokens are retained as $\mathbf{Q}_t = [\mathbf{q}_t^1, \ldots, \mathbf{q}_t^M] \in \mathbb{R}^{M \times d}$ ($M = 4$ by default). Each token can condition on preceding tokens, allowing their predictions to depend on one another. Their prospective role is learned through future-grounded ranking supervision, without assigning predefined future factors to individual tokens. At each rollout step, the tokens are regenerated from the currently available inputs. With $\theta = \{P_v, P_c, \psi, W_Q, W_K\}$, the prospective tokens serve as attention queries over eligible historical frames. We compute scaled dot-product attention logits and aggregate them with a smooth maximum:

$$a_{t,m,i} = \frac{(W_Q \mathbf{q}_t^m)^\top (W_K \mathbf{h}_i)}{\sqrt{d}}, \qquad \hat{r}_{t,i} = \tau_q \log\left(\frac{1}{M} \sum_{m=1}^{M} e^{a_{t,m,i}/\tau_q}\right). \tag{4}$$

Here $\tau_q > 0$ controls aggregation sharpness. A frame can score highly by matching any prospective token. The objective supervises the aggregated ranking without explicitly enforcing diversity among tokens or selected frames.

3.3 FUTURE-GROUNDED RANKING DISTILLATION

We train the prospective tokens by matching their predicted frame rankings to rankings derived from future content. Each training trajectory supplies eligible history, recent context, a known condition, and a realized continuation $\mathcal{Y}_t^+ = \{x_{t+j}\}_{j=1}^H$. The teacher compares candidate and continuation frames using a frozen DINOv2 encoder Φ. Historical views of subjects or scenes that recur in the continuation can provide reusable visual evidence, motivating the correspondence target

$$R^F_{t,i,j} = \cos(\Phi(x_{\tau_i}), \Phi(x_{t+j})), \qquad r^F_{t,i} = \tau_f \log\left(\frac{1}{H} \sum_{j=1}^{H} e^{R^F_{t,i,j}/\tau_f}\right). \tag{5}$$

This smooth maximum favors correspondence with at least part of the continuation. It provides a visual relevance proxy, rather than a measurement of incremental generation utility beyond the recent context or a guarantee of coverage across the continuation. We empirically examine the relationship between this proxy and downstream generation utility in Appendix E.5. The teacher uses only $(\mathcal{H}_t, \mathcal{Y}_t^+)$, while the student receives only $(\mathcal{H}_t, \mathcal{L}_t, c_t^+)$ during both training and inference.

Table 1: **VBench-Long results for 60-second generation on all 128 MovieGenBench prompts.** Each block compares long-context methods on one backbone. Average rank is computed over six metrics. Bold and underline mark the best and second-best results within each block.

Method	Subject ↑	Background ↑	Motion ↑	Dynamic ↑	Aesthetic ↑	Imaging ↑	Avg. Rank ↓
Self-Forcing	95.84	95.27	98.20	<u>51.72</u>	56.05	62.22	4.33
+∞-RoPE	97.24	96.24	98.58	46.64	56.09	63.28	3.17
+Deep Forcing	96.08	95.38	98.24	41.44	56.68	60.81	4.17
+LongLive-RAG	<u>97.60</u>	**96.51**	**98.70**	44.69	<u>57.19</u>	<u>64.97</u>	<u>2.00</u>
+FRAMEMORROW (Ours)	**97.71**	<u>96.45</u>	<u>98.60</u>	**64.56**	**57.27**	**68.47**	**1.33**
LongLive 1.0	97.13	95.89	98.61	44.56	58.17	67.56	3.83
+∞-RoPE	97.00	95.85	98.53	**53.36**	57.48	66.94	4.25
+Deep Forcing	<u>97.17</u>	96.04	<u>98.73</u>	45.13	57.48	67.27	3.42
+LongLive-RAG	**97.32**	<u>96.08</u>	98.62	49.90	<u>58.30</u>	<u>67.79</u>	<u>2.25</u>
+FRAMEMORROW (Ours)	**97.32**	**96.17**	**98.75**	<u>50.42</u>	**58.75**	**67.82**	**1.25**
Causal Forcing	93.52	94.12	95.74	72.32	51.24	62.30	4.83
+∞-RoPE	93.81	93.78	96.09	**92.47**	54.42	67.50	3.33
+Deep Forcing	94.27	94.18	**96.62**	78.59	52.12	64.25	3.17
+LongLive-RAG	<u>94.29</u>	<u>94.24</u>	96.48	88.20	<u>54.95</u>	<u>68.16</u>	<u>2.33</u>
+FRAMEMORROW (Ours)	**94.54**	**94.28**	<u>96.59</u>	<u>89.74</u>	**56.27**	**68.76**	**1.33**

We collect $\mathbf{r}_t^F = [r_{t,1}^F, \ldots, r_{t,N}^F]$ and normalize teacher and student scores into ranking distributions:

$$\mathbf{p}_t^F = \mathrm{Softmax}(\mathbf{r}_t^F/\tau_r), \qquad \hat{\mathbf{p}}_t = \mathrm{Softmax}(\hat{\mathbf{r}}_t/\tau_s). \tag{6}$$

The temperatures τ_f, τ_r, τ_s are positive. We retain pairwise orderings separated by a margin $\delta > 0$, $\mathcal{P}_t = \{(i,j) \mid r_{t,i}^F \geq r_{t,j}^F + \delta\}$, and optimize

$$\mathcal{L}_{\mathrm{sel}} = \mathcal{L}_{\mathrm{list}} + \lambda \mathcal{L}_{\mathrm{pair}}, \tag{7}$$

where the listwise term aligns the full distribution and the pairwise term preserves orderings separated by the teacher-score margin:

$$\mathcal{L}_{\mathrm{list}} = -\sum_{i=1}^{N} p_{t,i}^F \log \hat{p}_{t,i}, \qquad \mathcal{L}_{\mathrm{pair}} = \frac{1}{|\mathcal{P}_t|} \sum_{(i,j)\in\mathcal{P}_t} \mathrm{softplus}[-(\hat{r}_{t,i} - \hat{r}_{t,j})]. \tag{8}$$

We set $\mathcal{L}_{\mathrm{pair}} = 0$ when $\mathcal{P}_t$ is empty. At inference, continuation-based target construction is removed. The selector still encodes observed frames with E_v and predicts prospective tokens from $(\mathcal{H}_t, \mathcal{L}_t, c_t^+)$, without accessing $\mathcal{Y}_t^+$.

3.4 PLUG-AND-PLAY INTEGRATION

FRAMEMORROW outputs explicit historical frames, while each generator processes these frames through its own conditioning mechanism. For a frozen backbone b that supports reference-frame or memory conditioning, its native adapter Γ_b maps the selected frames to the representation already accepted by that model:

$$\mathbf{M}_t^{(b)} = \Gamma_b(\mathcal{M}_t), \qquad \hat{\mathcal{Y}}_t^+ = G_{\omega_b}(\mathcal{L}_t, c_t^+, \mathbf{M}_t^{(b)}), \qquad \omega_b \text{ is frozen.} \tag{9}$$

For example, a reference-frame interface applies its existing image preprocessing and reference encoder to $\mathcal{M}_t$. No new conditioning module is trained. The selector decides which historical frames to provide, while each generator retains its own way of processing the selected content.

Predicting a small number of prospective tokens avoids generating an additional future video for selection. A fixed K bounds the number of additional frames supplied to the generator. Selector processing still depends on the candidate count, recent-context length, and condition length.

4 EXPERIMENTS

We evaluate FRAMEMORROW on five benchmarks covering long-video generation, single-shot interactive generation, multi-shot interactive generation, closed-source generation, and action-conditioned world models.

Table 2: **Interactive video generation over 60 seconds.** Single-shot and multi-shot results are grouped separately. Bold marks the better result within each backbone pair.

	Overall			Segment-wise CLIP Score ↑						
Model	**Quality ↑**	**Consistency ↑**	**Aesthetic ↑**	**0–10s**	**10–20s**	**20–30s**	**30–40s**	**40–50s**	**50–60s**	**Avg.**
Single-shot										
LongLive 1.0	83.49	92.62	64.21	30.71	29.41	27.35	28.45	27.63	28.95	28.75
+FRAMEMORROW (Ours)	**84.06**	**93.21**	**64.56**	**30.72**	**29.65**	**28.11**	**28.78**	**28.03**	**29.18**	**29.08**
Self-Forcing	78.71	84.95	58.16	30.40	29.52	26.93	25.10	23.37	23.02	26.39
+FRAMEMORROW (Ours)	**81.74**	**89.30**	**60.62**	**30.56**	**29.67**	**28.34**	**27.12**	**26.60**	**26.01**	**28.05**
Causal Forcing	76.00	80.46	55.45	**29.92**	**27.37**	23.89	22.43	21.02	22.56	24.53
+FRAMEMORROW (Ours)	**78.66**	**83.48**	**58.75**	29.88	27.15	**24.27**	**23.76**	**23.30**	**23.08**	**25.24**
CausVid	81.94	89.77	63.53	**29.92**	29.00	28.34	28.22	28.19	28.08	28.63
+FRAMEMORROW (Ours)	**82.26**	**90.52**	**63.64**	29.81	**29.21**	**28.86**	**28.82**	**28.81**	**28.64**	**29.03**
Multi-shot										
LongLive 2.0	82.77	93.14	61.32	29.27	**28.12**	**28.12**	28.53	28.04	28.02	28.35
+FRAMEMORROW (Ours)	**83.18**	**93.49**	**61.68**	**29.28**	28.05	28.06	**28.71**	**28.29**	**28.24**	**28.44**
ShotStream	83.12	92.75	**62.30**	**30.15**	**28.87**	28.00	29.14	29.19	29.19	29.09
+FRAMEMORROW (Ours)	**84.85**	**96.76**	61.74	29.68	28.58	**28.38**	**29.36**	**29.43**	**29.51**	**29.16**

4.1 EXPERIMENTAL SETUP

Backbones and baselines. For long-video generation, we evaluate Self-Forcing (Huang et al., 2026), LongLive 1.0 (Yang et al., 2025), and Causal Forcing (Zhu et al., 2026). Long-context baselines include ∞-RoPE (Yesiltepe et al., 2026), Deep Forcing (Yi et al., 2025), and LongLive-RAG (Hu et al., 2026). Interactive video experiments additionally include CausVid (Yin et al., 2025), LongLive 2.0 (Chen et al., 2026b), and ShotStream (Luo et al., 2026b). World-model experiments use Matrix-Game 3.0 (Wang et al., 2026b), WorldMem (Xiao et al., 2026), and YuMe 1.5 (Mao et al., 2026). Closed-source model experiments use Seedance 2.0 and Kling O3 through their public reference-conditioning interfaces.

Offline training and integration. For text-conditioned generation, we train on 10K OpenVidHD Nan et al. (2025) videos organized into streaming prompts. For action-conditioned world models, we use 10K chunk-aligned examples from Sekai Game-Walking with pseudo-actions derived from camera motion. In both settings, realized future observations are used only to construct future-grounded ranking supervision. At inference, selected frames are passed through each frozen backbone's existing conditioning interface. Unless otherwise stated, we use $K = 4$ historical frames and $M = 4$ prospective tokens while retaining each backbone's native recent context. Further details are provided in Appendix A.

Evaluation metrics. For MovieGenBench, we report the six VBench-Long dimensions and average rank across these dimensions within each backbone. Ties receive their average rank. For interactive long video generation, we report overall quality, consistency, and aesthetic scores. We measure semantic adherence using CLIP scores between each 10-second segment and its corresponding prompt. For closed-source generation, we report the same overall quality, consistency, and aesthetic scores as in interactive video generation. For interactive world models, we report visual quality, temporal quality, and action alignment. Qwen3-VL-8B-Instruct evaluates action alignment by assessing whether generated rollouts follow the supplied actions.

4.2 MAIN RESULT

Long video generation. Table 1 compares long-context mechanisms for 60-second MovieGenBench generation. FRAMEMORROW achieves the best average rank on all three backbones: 1.33 for Self-Forcing, 1.25 for LongLive 1.0, and 1.33 for Causal Forcing. With Self-Forcing, it improves imaging quality from 62.22 to 68.47 and dynamic degree from 51.72 to 64.56. The advantage is not uniform across metrics, with LongLive-RAG retaining higher background and motion scores on Self-Forcing. The consistent average-rank advantage supports the effectiveness of frame selection across these generators, with trade-offs in individual metrics.

Interactive video generation. We evaluate interactive generation where the generation condition changes over time, covering both single- and multi-shot settings. Table 2 reports improvements in overall quality and consistency across both settings.

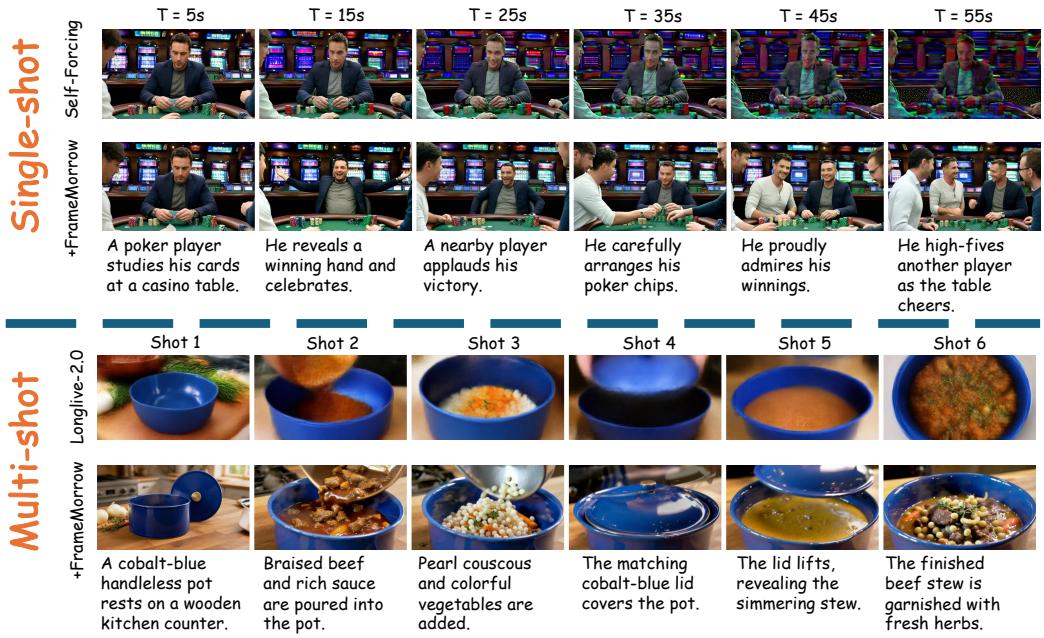

Figure 3: **Qualitative comparisons for interactive video generation.** Top: single-shot generation with Self-Forcing under sequential prompts. Bottom: multi-shot generation with LongLive 2.0. Each pair compares the backbone alone with its FRAMEMORROW-augmented variant.

For *single-shot generation*, FRAMEMORROW improves overall quality and consistency across all evaluated backbones. For Self-Forcing, consistency rises from 84.95 to 89.30. Its segment-wise CLIP-score gain increases from 0.16 in the first 10 seconds to 2.99 in the last 10 seconds, indicating larger improvements in prompt alignment at later stages of generation. Figure 3 shows a representative example, where the native Self-Forcing sequence develops strong color artifacts, while FRAMEMORROW depicts the successive actions at the poker table more clearly.

Table 3: **Closed-source generation.** All baselines use the same reference budget.

Model	Quality ↑	Consistency ↑	Aesthetic ↑
Seedance 2.0	86.80	94.02	65.61
+ Uniform	86.47	93.54	65.28
+FRAMEMORROW	**87.46**	**95.31**	**65.92**
Kling O3	83.47	91.28	63.41
+ Uniform	83.21	90.84	63.18
+FRAMEMORROW	**84.06**	**92.61**	**63.77**

For *multi-shot generation*, FRAMEMORROW also improves overall quality and consistency for both evaluated backbones. In particular, ShotStream gains 4.01 consistency points, although its aesthetic score decreases by 0.56. In the qualitative example, FRAMEMORROW better preserves the blue pot and the specified ingredients across multiple successive shot changes.

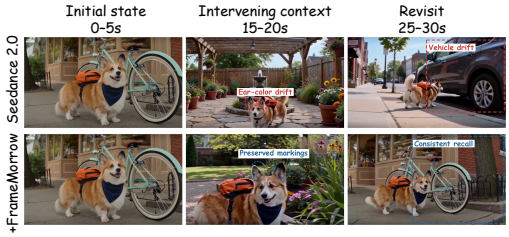

Figure 4: **Seedance 2.0 revisit case.**

Extension to closed-source generators. To test whether FRAMEMORROW remains effective without access to generator internals, we further evaluate it on Seedance 2.0 and Kling O3 through their public reference-conditioning interfaces. For all variants, the latest generated clip is provided through the native video-reference interface. Uniform-K and FRAMEMORROW additionally receive the same number of historical frames through the image-reference interface, differing only in how these frames are selected. Table 3 reports 30 cases 30-second multi-turn generation results using the same overall metrics as our open-source evaluation. Compared with the original generators, FRAMEMORROW improves consistency by 1.29 points on Seedance 2.0 and 1.33 points on Kling O3. It also outperforms uniform selection on all three reported metrics for both models. These results support plug-and-play use with closed-source generators and show the benefit of selecting relevant references under a fixed reference budget. Figure 4 shows a representative revisit case on Seedance 2.0. While the original model exhibits subject and

Table 4: **Action-conditioned interactive world generation.** Subject, background, anti-flicker, motion, and action-alignment scores lie in $[0, 1]$. Imaging quality is on a 0–100 scale. Shaded rows denote FRAMEMORROW, and bold indicates the better result within each base-model pair.

	Visual Quality			Temporal Quality		Interaction
Model	**Subject ↑**	**Background ↑**	**Imaging ↑**	**Anti-flicker ↑**	**Motion ↑**	**Action Alignment ↑**
Matrix-Game 3.0	0.801	0.894	68.20	0.936	**0.950**	0.842
+FRAMEMORROW (Ours)	**0.816**	**0.910**	**68.35**	**0.939**	0.949	**0.861**
WorldMem	0.782	0.924	70.23	0.910	0.916	0.866
+FRAMEMORROW (Ours)	**0.793**	**0.932**	**70.47**	**0.913**	**0.918**	**0.889**
YuMe 1.5	0.765	0.872	50.98	0.944	0.965	0.883
+FRAMEMORROW (Ours)	**0.798**	**0.886**	**51.80**	**0.947**	**0.967**	**0.912**

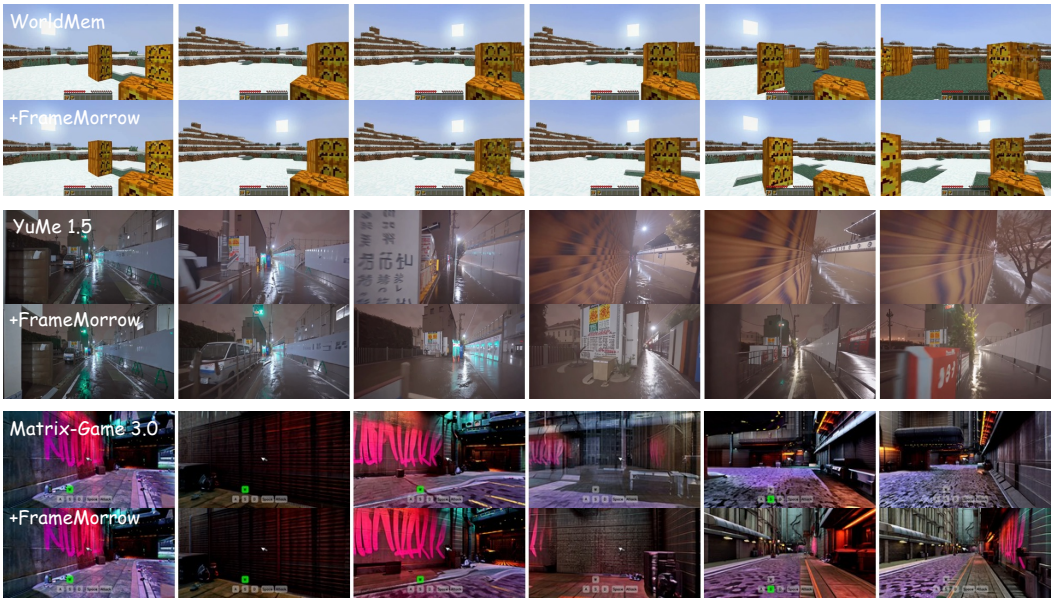

Figure 5: **Qualitative comparisons for interactive world generation.** Each pair shows the original backbone (top) and +FRAMEMORROW (bottom) on WorldMem, YuMe 1.5, and Matrix-Game 3.0.

object drift after intervening interactions, FRAMEMORROW better preserves both the established subject appearance and the previously observed vehicle.

Action-conditioned interactive world models. Table 4 shows consistent improvements across the evaluated world models. FRAMEMORROW improves subject and background consistency, imaging quality, anti-flicker, and action alignment on all three backbones. Action alignment increases by 0.019 for Matrix-Game 3.0, 0.023 for WorldMem, and 0.029 for YuMe 1.5, while motion quality is largely preserved. These results demonstrate that FRAMEMORROW strengthens both long-range visual consistency and control alignment across diverse interactive world models. Figure 5 provides corresponding visual examples. WorldMem better preserves the snow-covered ground, while YuMe 1.5 and Matrix-Game 3.0 retain more consistent street and corridor structures over longer rollouts.

4.3 ANALYSIS AND ABLATIONS

We analyze three design choices on the 60-second single-shot interactive generation setting with a frozen Self-Forcing backbone. The recent-context window retains the backbone's native default, and the eligible-history rule is fixed. Only older historical frames count toward K.

Future supervision. Figure 6(a) compares recent-context supervision (the no-future control), future distillation, and a future-informed oracle. The no-future control constructs ranking targets from DINOv2 similarities between eligible historical frames and recent-context observations, whereas future distillation uses realized future observations. Their gains over the native score of 84.95 are 1.25, 4.35, and 5.65 points. Future distillation adds 3.10 points over the control and reduces the

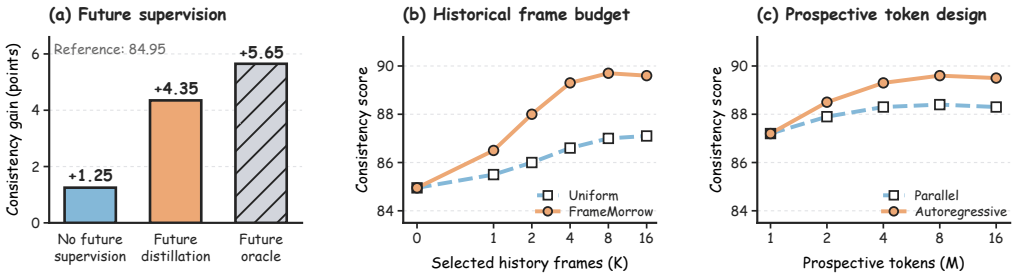

Figure 6: **Ablation studies.** Future supervision, frame budget, and token design.

oracle gap from 4.40 to 1.30 points, supporting continuation-derived targets for causal selection. The oracle accesses reference future observations and serves as a diagnostic comparison.

Historical frame budget. Figure 6(b) varies the number of selected historical frames K while fixing $M = 4$, and compares FRAMEMORROW with uniform sampling from the same eligible history. FRAMEMORROW reaches 88.00 with only two selected frames, exceeding uniform sampling with sixteen frames (87.10). Increasing K from 4 to 8 adds only 0.40 points, with no further improvement at $K = 16$. These results show that selecting relevant historical frames matters more than simply increasing their number.

Prospective token design. Figure 6(c) compares autoregressive and parallel prediction of prospective tokens at $K = 4$. Increasing the number of autoregressively predicted tokens from $M = 1$ to $M = 4$ raises consistency from 87.20 to 89.30. At $M = 4$, autoregressive prediction exceeds parallel prediction by 1.00 point. Increasing M to 8 adds only 0.30 points, with no further gain at $M = 16$. Four autoregressive tokens capture most of the improvement.

Historical frame selection. Table 5 compares consistency for 60-second single-shot generation with Self-Forcing and multi-shot generation with LongLive 2.0. All strategies share the backbone, eligible history, memory budget, and conditioning interface. *1) Recent-K* selects the latest eligible frames. *2) Uniform* samples frames uniformly from the eligible history. *3) Context matching* ranks frames by cosine similarity to the mean recent-context feature from the selector's frozen visual encoder. *4) Prompt matching* ranks frames by frozen CLIP image–text similarity to the next-shot or next-segment prompt. *5) Direct scoring* uses the same inputs to predict each candidate's score from its frame without prospective tokens. FRAMEMORROW outperforms the five selection strategies.

Table 5: **Historical frame selection.**

Strategy	Single-shot	Multi-shot
Recent-K	85.70	92.80
Uniform	86.60	92.21
Context matching	87.85	92.68
Prompt matching	88.35	93.17
Direct scoring	88.22	92.53
FRAMEMORROW	**89.30**	**93.49**

Inference overhead. We measure the inference overhead of FRAMEMORROW on three models under matched hardware and sampling settings (Table 6). For all backbones, FRAMEMORROW uses $K = 4$ and $M = 4$, with selection refreshed every 5 seconds. Selection time includes feature encoding, prospective token prediction, scoring, and Top-K per refresh. Total time covers the pipeline, including historical-frame conditioning, and Δ denotes the increase over the native backbone.

Table 6: **Inference overhead.**

Method	Sel. (ms)	Total (s)	Δ (%)
Self-Forcing	–	82.6	–
+FRAMEMORROW	112	84.0	1.7
LongLive 1.0	–	67.8	–
+FRAMEMORROW	109	69.2	2.1
Causal Forcing	–	82.6	–
+FRAMEMORROW	115	84.0	1.7

5 CONCLUSION

We present FRAMEMORROW, which uses *prospective tokens* to predict future information needs and select relevant historical frames. Its explicit-frame interface enables lightweight, plug-and-play use across diverse generators. Experiments on five benchmarks and 11 models demonstrate improvements in long-range consistency, visual quality, and action alignment. These results support selecting historical information according to future needs rather than current relevance alone.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and code assistance. We reviewed and revised all AI-assisted text, and manually verified and tested AI-assisted code for correctness. Generative AI was not used for research ideation, experimental design, result analysis, or scientific claim generation. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16173–16184, 2026.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29071–29082. IEEE, 2025.

Hanmo Chen, Chenghao Xu, Xu Yang, Xuan Chen, and Cheng Deng. Past-and future-informed kv cache policy with salience estimation in autoregressive video diffusion. *arXiv preprint arXiv:2601.21896*, 2026a.

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

Yixuan Ding, Jiahao Kong, Wei Huang, Ruijie Quan, and Yi Yang. Layerrecall: A state-conditioned memory router for long-horizon consistency in video generation. *arXiv preprint arXiv:2608.28460*, 2026.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712. IEEE, 2025.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

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

Chaoyu Li, Tianzhi Li, Fei Tao, Zhenyu Zhao, Ziqian Wu, Maozheng Zhao, Juntong Song, Cheng Niu, and Pooyan Fazli. Frameoracle: Learning what to see and how much to see in videos. *arXiv preprint arXiv:2510.03584*, 2025.

Jiayi Luo, Qiyan Liu, Tengyang Wang, JunHao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, Qingyun Sun, et al. Future forcing: Future-aware training-free kv cache policy for autoregressive video generation. *arXiv preprint arXiv:2605.30083*, 2026a.

Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling. *arXiv preprint arXiv:2603.25746*, 2026b.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, volume 2025, pp. 1045–1064, 2025.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16944–16953, 2026.

Hongjun Wang, Lin Liu, Jianguo Li, and Tao Lin. Dual-granularity memory for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38016–38026, 2026a.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Xuan Wei, Longbin Ji, Guan Wang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Qingqi Hong. Memento: Reconstruct to remember for consistent long video generation. *arXiv preprint arXiv:2606.14667*, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

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

Bo Ye, Xinyu Cui, Jian Zhao, Tong Wei, and Min-Ling Zhang. Dysink: Dynamic frame sinks for autoregressive long video generation. *arXiv preprint arXiv:2605.21028*, 2026.

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.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, volume 2025, pp. 84154–84179, 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.

Liuzhou Zhang, Jiarui Ye, Yuanlei Wang, Ming Zhong, Mingju Cao, Wanke Xia, Bowen Zeng, Zeyu Zhang, and Hao Tang. Egolcd: Egocentric video generation with long context diffusion. *arXiv preprint arXiv:2512.04515*, 2025b.

Zirui Zheng, Jiaqian Yu, Xiongfeng Peng, Mingyi Li, Chao Zhang, Weiming Li, Dong Wang, Huchuan Lu, Xu Jia, et al. Mem-world: Memory-augmented action-conditioned world models for persistent robot manipulation. *arXiv preprint arXiv:2606.18960*, 2026.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Implementation Details

A.1 Selector Architecture

The selector takes eligible history, recent context, and the known rollout condition. Visual and condition encoders are frozen. Only the visual/condition projections, causal Transformer, and ranking query/key projections are trained, without the generator. Compatible text-conditioned backbones share one checkpoint. Action-conditioned experiments use a separate checkpoint with the same architecture.

Table 7 summarizes the architecture. DINOv2 ViT-B/14 provides final class-token frame features for the visual encoder and teacher. Learned projections map visual and condition features into the selector space. Text conditions use frozen UMT5-XXL token features. Actions are serialized as short motion descriptions and encoded by frozen CLIP ViT-B/32, with a separate projection for its feature dimension.

Historical and recent tokens follow observation time, with learned temporal and token-type embeddings distinguishing history, recent context, and conditions. A learned start token initializes autoregressive queries under a causal mask. Padding is masked in attention and ranking. Only eligible history is scored and returned by Top-K, excluding recent context. Deployment retains the backbone's native recent window. If fewer than K candidates exist, all are returned without duplication.

Table 7: **Selector architecture.** Frozen encoders are excluded from the trainable parameter estimate.

Configuration	Value
Visual encoder / teacher	DINOv2 ViT-B/14
Visual feature dimension	768
Transformer layers / hidden dimension	4 / 512
Attention heads / FFN dimension	8 / 2,048
Normalization / activation	Pre-LayerNorm / GELU
Dropout	0.0
Positional encoding	Learned temporal embeddings
Prospective queries M / memory budget K	4 / 4
Maximum training candidates	128
Trainable parameters, text selector	16M

A.2 Ranking Targets and Losses

The teacher computes cosine similarities between ℓ_2-normalized DINOv2 features of historical and realized future frames. Future observations supply detached targets only and never enter the selector. Scores are aggregated by the smooth maximum in Section 3.3, then softmaxed over valid historical candidates for listwise supervision, without per-example min–max normalization.

The teacher uses $H = 8$ continuation frames sampled at two fps over the next four seconds. Table 8 lists the loss settings. Pairwise softplus loss is averaged over all ordered pairs with teacher-score gaps of at least δ, or set to zero if none qualify. The listwise term remains active. Both losses are averaged across examples, not pooled candidates.

A.3 Optimization and Compute

Text- and action-conditioned selectors are trained independently using precomputed frozen encoder features. Table 9 gives the configuration. Linear warmup precedes cosine learning-rate decay, gradients are clipped before each update, and validation listwise loss selects the final checkpoint. The planning budget is six hours on two A100 80GB GPUs, or 12 GPU-hours per selector, excluding annotation and feature extraction.

Table 8: **Ranking hyperparameters.**

Symbol	Role	Value
τ_q	Query-score aggregation	0.10
τ_f	Future-similarity aggregation	0.05
τ_r	Teacher ranking temperature	0.10
τ_s	Student ranking temperature	1.00
δ	Teacher margin for pair inclusion	0.05
λ	Pairwise loss weight	0.50
H	Number of teacher continuation frames	8

Table 9: **Optimization settings.**

Configuration	Value
Optimizer	AdamW
Learning rate / weight decay	10^{-4} / 0.01
Adam coefficients	$(0.9, 0.999)$
Schedule / warmup	Cosine / 5% of updates
Global batch size / epochs	64 / 5
Precision / gradient clipping	BF16 / global norm 1.0
Training seed	0
Hardware budget	2× A100 80GB

B Training Data Construction

B.1 Text-conditioned Training Tuples

We use 10K OpenVidHD videos, divided into two-second annotation units with four uniformly sampled frames each. These units do not set rollout-condition durations or selection refresh intervals. Qwen3-VL-8B-Instruct generates structured segment descriptions, which form the streaming conditions.

Each tuple $(\mathcal{H}_t, \mathcal{L}_t, c_t^+, \mathcal{Y}_t^+)$ contains eligible history, recent context ending at t, the next-rollout condition, and a teacher-only continuation. Recent context comprises eight frames from the preceding four seconds. Up to 128 historical frames are sampled uniformly before this window and kept in temporal order. Boundaries lacking eligible history or a complete continuation are skipped. This training window does not replace the backbone's native deployment window.

Videos are split 90:10 for training and validation before tuple extraction, keeping all tuples from each video together. Teacher frames remain within the condition's temporal scope. Shorter intervals use fewer frames rather than sampling beyond that scope.

B.2 Segment Annotation and Prompt Construction

The annotation template preserves visible identities, scene attributes, actions, and state changes without introducing unsupported objects or events.

Prompt 1: Segment annotation

Instructions
Given four chronologically ordered frames from a two-second video segment, describe the visible subjects, scene, actions, and object states. If a preceding segment description is supplied, identify visible changes while preserving consistent names for recurring subjects and objects. Do not infer unseen events or future actions.

Output
Return the fields `subject`, `scene`, `action`, `state`, `change`, and `segment_prompt`. The `segment_prompt` should be a concise visual description suitable for conditioning this segment.

Table 10: **Backbone integration map.** Every row retains the native recent window and selects up to four additional historical frames. No generator fine-tuning is used.

Backbone	Selector	Historical conditioning route
Self-Forcing	Text	Frame encoding followed by historical KV conditioning
LongLive 1.0	Text	Historical frame encoding into the memory cache
Causal Forcing	Text	Frame encoding into causal context KV states
CausVid	Text	Historical latent encoding into context KV states
LongLive 2.0	Text	Historical references in the streaming memory interface
ShotStream	Text	Selected images as cross-shot reference conditioning
Matrix-Game 3.0	Action	Encoded historical frames in the rollout context
WorldMem	Action	Historical frame memory with associated pose metadata
YuMe 1.5	Action	Historical latent references for rollout conditioning

Structured fields form each segment prompt, excluding later timestamps and teacher scores. Annotation uses temperature zero and at most 256 output tokens. Empty or malformed outputs are regenerated once, then excluded if still invalid.

B.3 Action-conditioned Training Tuples

We use 10K chunk-aligned Sekai Game-Walking examples. Each combines observed history and context, the next chunk's pseudo-action condition, and its realized continuation for the teacher. Pseudo-actions derive from relative camera translation and rotation and need not reproduce the original keyboard inputs.

Under the camera-to-world pose convention, let R_t and $\mathbf{p}_t$ denote camera orientation and position. Local translation and angular velocity are computed as

$$\mathbf{v}_t = R_t^\top(\mathbf{p}_{t+1} - \mathbf{p}_t)/\Delta t, \qquad \boldsymbol{\omega}_t = \mathrm{Log}(R_t^\top R_{t+1})^\vee/\Delta t. \tag{10}$$

Translation is normalized by the trajectory's median nonzero speed, with components activated above 0.2. Yaw and pitch use a 2°/s threshold. Labels cover forward/backward, left/right, turn left/right, and look up/down. Simultaneous labels are retained, and chunks are stationary when all components fall below threshold. A trajectory-level 90:10 train/validation split keeps adjacent chunks together.

C Backbone Integration Details

C.1 Shared Selection and Backbone-specific Conditioning

The selector returns explicit frames and shares one checkpoint per condition modality across compatible generators, with $K = M = 4$. Each backbone retains its native recent window. Integration trains neither the generator nor a new conditioning module.

Selected frames are sorted by timestamp and converted by each backbone's adapter into its reference representation (Table 10). Generator activations never feed back into the selector.

C.2 Temporal Positions and Cache Management

For KV conditioning, the backbone's frozen image/video encoder and context-encoding path convert selected frames into layer-specific keys and values. DINOv2 features are not copied into this cache. Reference latents are conditioning inputs, not generated output frames. Backbones with direct image-reference interfaces instead receive selected RGB frames.

Selection retains original timestamps. Adapters use historical temporal positions where supported, or map sorted frames to reference slots in temporal order. Refreshes replace historical slots while preserving native recent context, whose frames are excluded from the candidate pool. Thus, the four-frame budget fixes additional observations, not latent-token or KV-cache size across backbones.

C.3 Matched Integration Controls

Recent-K, Uniform, Context matching, Prompt matching, and FrameMorrow share eligible history, budget, and conditioning adapter within each backbone. Context matching uses the selector's recent visual features, whereas Prompt matching uses CLIP image–text similarity to the known condition. These controls isolate selection quality. Comparisons with the unaugmented backbone also include the benefit of adding historical references.

D Evaluation Protocols

Unless otherwise specified, historical frame selection is refreshed every five seconds of generated video. Each refresh uses only the eligible history, recent context, and condition available at that time.

D.1 Long Video Generation

We generate 60-second videos from 128 MovieGenBench prompts and evaluate six VBench-Long dimensions: subject/background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. They assess foreground/scene persistence, temporal smoothness, motion magnitude, aesthetics, and frame-level quality, respectively. Higher is better for all six.

Prompt $i \in \{0, \ldots, 127\}$ uses paired seed i. Each backbone and its augmented variants share its released resolution, frame rate, denoising steps, and guidance. Settings are matched within, not across, backbones. Evaluation covers the full 60 seconds without selecting favorable subsequences.

Methods are ranked per metric within each backbone before display rounding, with ties assigned average ranks. Average rank is the mean over six metrics. Dataset scores weight prompts equally.

D.2 Interactive Video Generation

We evaluate 100 single-shot and 100 multi-shot cases, each comprising six ten-second condition intervals. LLM-generated prompt sequences require dependencies across intervals. Single-shot cases retain one scene, while multi-shot transitions depend on previously established subjects and states. Methods share prompts and recent context. Selection refreshes every five seconds, independently of the ten-second condition/evaluation intervals.

Segment-wise alignment pairs each ten-second interval with its current prompt, following the evaluation organization used for LongLive. In the CLIP ViT-B/32 implementation, eight uniformly sampled frames per interval are encoded and compared with the normalized prompt embedding. The score for interval j is

$$s_j = \frac{100}{n_j} \sum_{u=1}^{n_j} \cos\big(E^{\mathrm{img}}_{\mathrm{CLIP}}(x_{j,u}), E^{\mathrm{txt}}_{\mathrm{CLIP}}(c_j)\big). \tag{11}$$

Each interval's score is averaged across cases using only its current prompt, never prompts from later intervals.

Quality, Consistency, and Aesthetic use the VBench quality aggregate, subject consistency, and aesthetic evaluator, respectively. Multi-shot evaluation measures recurring-subject consistency across shots and temporal quality within shots, avoiding penalties for intended scene changes.

D.3 Closed-Source Generation

We evaluate Seedance 2.0 and Kling O3 through their public reference-conditioning interfaces via fal.ai on 30 cases with five turns each. Methods share prompts, initial context, explicit API seeds, and reference budgets. The native model receives the latest clip as a video reference. Uniform-K and FrameMorrow additionally receive equal numbers of historical image references, differing only in frame selection. Other settings remain fixed within each backbone.

Table 11: **Action-alignment rubric.**

Score	Criterion
1.00	All commanded motion directions and their temporal order are clearly followed.
0.75	The dominant commands are followed, with minor delay or one brief inconsistency.
0.50	Some commands are followed, but a substantial part is missing or ambiguous.
0.25	Only weak evidence supports the commands, with mostly inconsistent motion.
0.00	The commanded motion is absent, reversed, or unsupported by the visible rollout.

Table 12: **Sensitivity to the visual teacher.** All variants use the same selector architecture and training data. Δ denotes the improvement over the native Self-Forcing consistency score of 84.95.

Visual Teacher	Consistency ↑	Δ
CLIP	88.42	+3.47
SigLIP	88.76	+3.81
DINOv2	**89.30**	**+4.35**

D.4 Action-conditioned World Models

We evaluate Matrix-Game 3.0, WorldMem, and YuMe 1.5 on 60 trajectories each. Native and FrameMorrow variants share initial context and actions. Visual metrics cover subject/background consistency, anti-flicker, motion smoothness, and imaging quality. Qwen3-VL-8B-Instruct judges whether rollouts follow the supplied controls.

The judge receives 16 ordered frames per action interval, command descriptions, and durations, without method identities or competing outputs. Each interval is scored once at temperature zero using Table 11. Scores are averaged within trajectories and then across trajectories, avoiding extra weight for longer sequences.

Prompt 2: Action-alignment evaluation

Instructions
Compare camera motion with the supplied controls and intervals. Distinguish camera motion from moving objects. Judge action compliance, not aesthetics. Using the rubric, assign 0, 0.25, 0.5, 0.75, or 1. Do not infer invisible motion.

Inputs
Controls: `<action sequence and intervals>`
Frames: `<chronologically sampled rollout frames>`

Output
Return JSON fields `score` and `reason`, citing visible evidence.

Malformed outputs are retried once, with persistent failures reported as missing. The judge assesses observable action compliance rather than physical correctness.

E Additional Ablations

Unless otherwise specified, additional ablations use 60-second single-shot interactive generation with frozen Self-Forcing, its native recent window, $K = 4$ historical frames, and $M = 4$ prospective queries. We report overall consistency, as in the main ablations.

E.1 Sensitivity to the Visual Teacher

Visual teacher. We replace the DINOv2 teacher with CLIP or SigLIP, fixing selector architecture, training trajectories, candidate history, and optimization. All three improve consistency (Table 12), supporting the use of different visual feature spaces. DINOv2 performs best and is used in the main experiments.

Table 13: **Ablation of the ranking objective.** The full objective combines listwise distribution matching with margin-based pairwise ordering constraints.

Objective	Consistency ↑	Δ
Pairwise only	88.21	+3.26
Listwise only	88.67	+3.72
Listwise + Pairwise	**89.30**	**+4.35**

Table 14: **Effect of the realized-continuation horizon.** H denotes the number of continuation observations available only to the training teacher.

H	Consistency ↑	Δ
1	87.86	+2.91
2	88.71	+3.76
4	89.14	+4.19
8	**89.30**	**+4.35**

E.2 Ranking Objective

Listwise and pairwise supervision. Listwise loss transfers the teacher's relevance distribution, while pairwise loss preserves orderings above the teacher-score margin. Table 13 compares each alone with their combination. Both help, listwise supervision is stronger alone, and adding pairwise constraints yields the best consistency.

E.3 Length of the Realized Continuation

Continuation horizon. We vary the number of teacher continuation observations H, which affects supervision but not deployed inputs or computation. Consistency improves substantially up to $H = 4$, then largely saturates (Table 14). Multiple future observations provide a more reliable target than one frame, with limited benefit from longer horizons.

E.4 Aggregation of Prospective Queries

Query-score aggregation. We compare mean, hard-maximum, and smooth-maximum pooling of query–frame scores. Mean pooling favors consistent matches across queries, while hard maximum keeps only the strongest match. Smooth maximum allows one strong prospective match to dominate while retaining other queries' contributions and gives the best consistency (Table 15).

E.5 Relation Between Future Relevance and Generation Utility

Generation-utility validation. The target in Eq. (5) measures visual correspondence with the realized continuation, which need not imply downstream benefit. We test whether teacher relevance $r^F_{t,i}$ correlates with the generation utility of individual historical frames.

This diagnostic uses condition-free long-video generation. All ranking signals exclude upcoming conditions. At each rollout step t, we sample N_u eligible candidates $x_{\tau_i} \in \mathcal{H}_t$.

The privileged teacher score $r^F_{t,i}$ uses the held-out realized continuation, not generated futures. Generated continuations are used only to measure each candidate's downstream utility.

We first generate a baseline continuation using only the backbone's native recent context,

$$\hat{\mathcal{Y}}^0_t = G_\omega(\mathcal{L}_t), \tag{12}$$

and then generate an additional continuation for each candidate while providing that candidate through the same historical-conditioning interface,

$$\hat{\mathcal{Y}}^i_t = G_\omega(\mathcal{L}_t, x_{\tau_i}). \tag{13}$$

All generations share the initial context, seed, sampling settings, and frozen backbone. Only the supplied historical frame differs.

Table 15: **Aggregation across prospective queries.** All variants use $M = 4$ queries and the same trained selector configuration.

Aggregation	Consistency ↑	Δ
Mean	88.43	+3.48
Hard maximum	88.91	+3.96
Smooth maximum	**89.30**	**+4.35**

Table 16: **Relationship between historical-frame ranking signals and empirical generation utility.** Spearman correlations are computed across historical candidates at each rollout step, averaged within each trajectory, and then averaged across the held-out evaluation set. The future-grounded teacher uses the held-out realized continuation only for this diagnostic analysis, whereas FRAMEMORROW does not observe future content at inference. Confidence intervals are obtained by trajectory-level bootstrap with 1,000 resamples.

Ranking signal	Realized future	ρ ↑	95% CI
Context similarity	No	0.10	[0.04, 0.16]
FRAMEMORROW prediction $\hat{r}$	No	0.42	[0.35, 0.49]
Future-grounded teacher r^F	Yes	0.51	[0.44, 0.57]

We define the empirical generation utility of candidate i as

$$u_{t,i} = Q\left(\hat{\mathcal{Y}}_t^i\right) - Q\left(\hat{\mathcal{Y}}_t^0\right), \tag{14}$$

where Q is the main experiments' long-range consistency evaluator. It measures preservation of subjects, objects, and scene appearance from preceding context, rather than internal consistency of the new continuation alone. Positive utility indicates better preservation than the native backbone.

Because trajectory difficulty and utility scale may vary across examples, we measure rank correspondence among historical candidates at each rollout step rather than pooling raw utility values globally. Specifically, we compute

$$\rho_t(s, u) = \text{Spearman}\left(\{s_{t,i}\}_{i=1}^{N_u}, \{u_{t,i}\}_{i=1}^{N_u}\right), \tag{15}$$

where s denotes a historical-frame ranking signal. We compare current-context similarity, the privileged future-grounded teacher score r^F, and the deployed FRAMEMORROW prediction $\hat{r}$.

We average valid step-level correlations within each trajectory, then across trajectories. The analysis includes $N_{\text{traj}} = 30$ held-out trajectories, $N_{\text{step}} = 120$ valid steps, and $N_u = 8$ candidates per step. We obtain 95% confidence intervals by resampling trajectories with replacement 1,000 times.

Table 16 shows weak correspondence between current-context similarity and generation utility. The teacher ranking aligns more closely with utility, and FRAMEMORROW recovers much of this association without future access at inference. These results support continuation-based visual correspondence as a training proxy for downstream usefulness.

F Additional Qualitative Results

Figures 7, 8, and 9 compare native and FRAMEMORROW-augmented Self-Forcing, LongLive, and Causal Forcing under matched generation conditions and sampling settings, illustrating long-horizon visual consistency.

G User Study

We conduct a blinded pairwise study of long-video generation. For each of Self-Forcing, LongLive, and Causal Forcing, 30 randomly sampled native/FRAMEMORROW video pairs share prompts and generation settings. All ten raters evaluate every pair, yielding 300 judgments per backbone and 900 per criterion. Method identities are hidden and left–right order is randomized. Raters independently assess *long-range consistency* (preservation of subjects, objects, and scenes) and *overall*



Figure 7: **Additional qualitative comparisons with Self-Forcing.** Each example compares the original Self-Forcing model (top) with Self-Forcing+FRAMEMORROW (bottom) over long-horizon generation. FRAMEMORROW better preserves subject appearance and visual details as generation progresses.

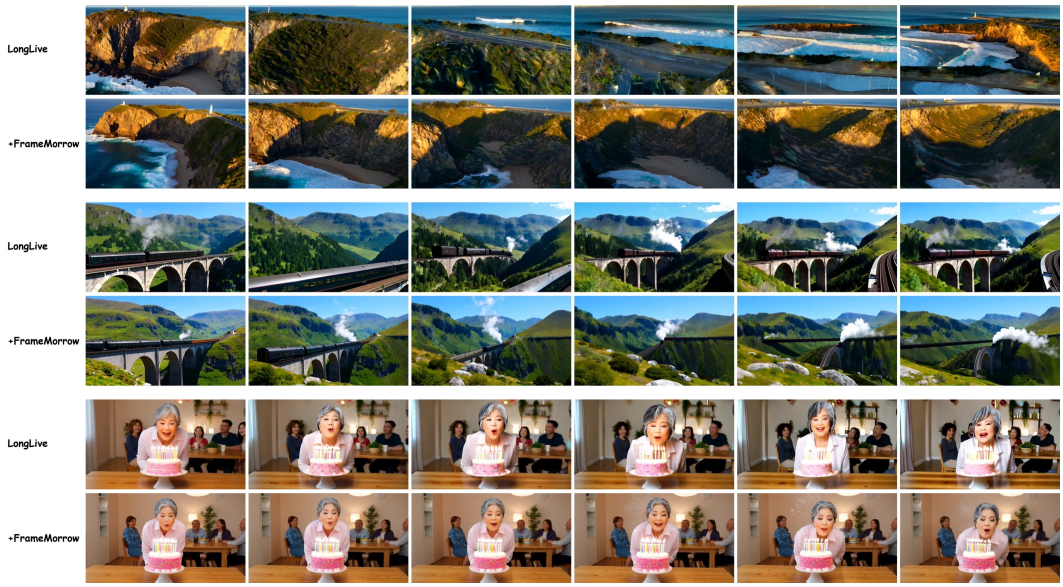

Figure 8: **Additional qualitative comparisons with LongLive.** Each example compares the original LongLive model (top) with LongLive+FRAMEMORROW (bottom) under the same generation conditions. FRAMEMORROW improves the preservation of subjects and scene appearance over extended generation.

visual quality (clarity, artifacts, and naturalness), selecting a preferred video or a tie when neither is clearly better.

Figure 10 reports preferences, with Overall averaging the three backbones. FRAMEMORROW is preferred more often than native models for both criteria on every backbone. Average consistency preferences are 62.7% for FRAMEMORROW, 16.7% for native, and 20.7% ties. Quality preferences

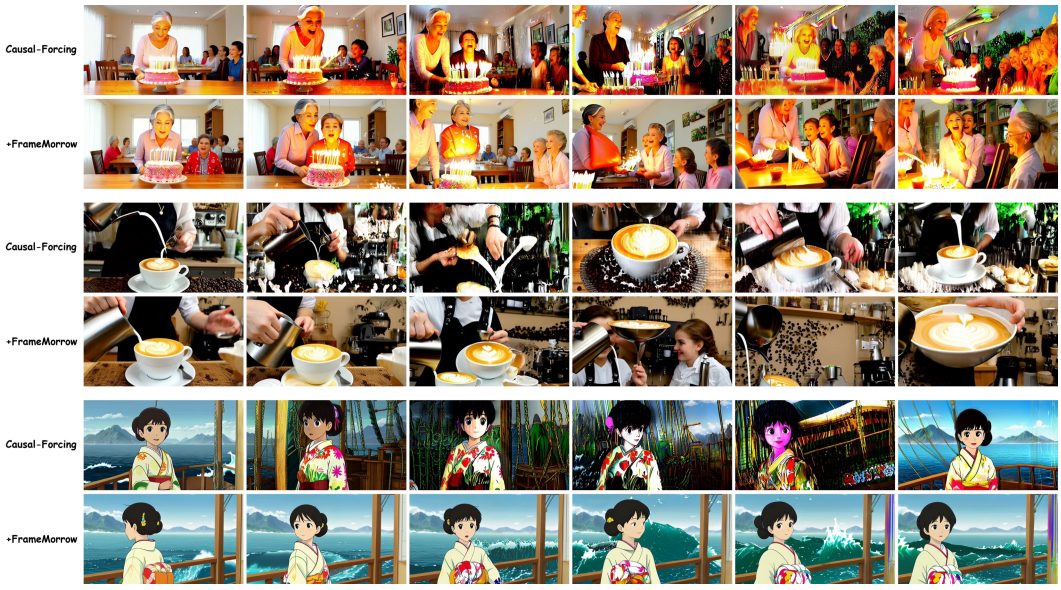

Figure 9: **Additional qualitative comparisons with Causal Forcing.** Each example compares the original Causal Forcing model (top) with Causal Forcing+FRAMEMORROW (bottom) under the same generation conditions. FRAMEMORROW provides more consistent subjects, objects, and scene details throughout long-horizon generation.

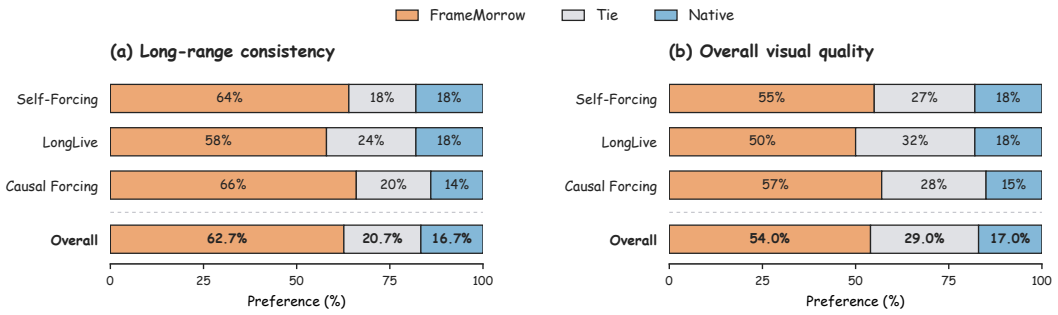

Figure 10: **User preferences for long-video generation.** FRAMEMORROW is preferred over the native model across all three backbones. Overall averages the three backbone-level percentages.

are 54.0%, 17.0%, and 29.0%, respectively. The stronger consistency preference matches our focus on historical preservation, while quality preferences indicate broader perceptual benefits.