ARE WE CACHING THE RIGHT HISTORY FOR AUTOREGRESSIVE VIDEO GENERATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

In autoregressive video generation, the historical key-value (KV) cache is crucial for long-term visual consistency, yet GPU memory and real-time inference constraints require it to remain bounded. A more recent view decomposes historical KV into sink, middle, and recent regions, yet prior studies have centered more on the two boundary regions, leaving the expanding and potentially information-rich middle KV largely unexplored. We systematically investigate middle KV across diverse autoregressive video backbones from two perspectives: (1) its functional role, showing that retaining middle KV improves long-range visual consistency and that it serves as a visual reference rather than an event memory; and (2) its underlying attention mechanism, revealing broadly distributed attention across the historical context rather than concentration on a few frames or prompt-transition boundaries. Based on these findings, we propose Temporal-Weighted Reservoir (TWR), a training-free cache policy that combines reservoir sampling with a time-weighted survival rule. TWR maintains broad coverage of an unknown-length middle history while preferentially preserving earlier visual references, improving long-horizon consistency under matched cache budgets. Extensive evaluations across diverse backbones consistently validate both the effectiveness of TWR and our functional analysis of middle KV.

1 INTRODUCTION

Recent advances in autoregressive video generation have made real-time streaming increasingly feasible, enabling interactive long-horizon applications such as controllable visual worlds and game-like simulations (Yin et al., 2025; Huang et al., 2025a; Yang et al., 2026; Liu et al., 2026). In autoregressive rollout, the historical key-value (KV) cache plays a critical role in maintaining long-term visual consistency, including subject identity, appearance, scene layout, and motion continuity. However, retaining the full history causes the KV cache to grow continuously with the rollout, leading to increasing memory and computational overhead. In practice, GPU memory constraints and real-time inference requirements therefore necessitate a bounded KV cache.

Existing bounded-cache strategies are largely motivated by practical effectiveness, typically retaining historical states through a recent sliding window or a sink-plus-recent layout (Huang et al., 2025a; Yang et al., 2026; Liu et al., 2026), or through content-based top-k selection (Yi et al., 2026a; Zhang et al., 2025). While these mechanisms provide efficient ways to control cache growth, they offer limited understanding of which parts of the historical KV cache are actually responsible for different aspects of generation. More recently, a three-part view that decomposes the historical cache into sink, middle, and recent regions has increasingly emerged (Yi et al., 2026a; Chen et al., 2026b). Under this decomposition, prior studies have centered more on the two boundary regions, showing that sink KV can serve as a persistent anchor, while recent KV provides the local context necessary for short-term temporal continuity. However, the functional role of middle KV remains largely unexplored. We argue that this region is equally important, because it spans an increasingly large portion of the generation history as autoregressive rollout proceeds, accumulating a progressively richer record of past visual states and thus potentially carrying substantial long-range information. Understanding how such information contributes to generation is therefore critical for principled KV-cache design in long-horizon video generation.

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.

To this end, we systematically investigate the functional nature of middle KV across diverse autoregressive backbones progressively. We begin by asking: *(1) Does retaining middle KV benefit generation?* By comparing generation with and without middle KV while keeping sink and recent KV fixed, we find that middle KV has a clear causal effect on long-range visual consistency. Building on this finding, *(2) what information does middle KV provide?* We test whether the model retrieves matched past events and whether controlled identity or structural changes introduced into middle KV propagate to later generated frames. Our results show that middle KV serves primarily as a visual reference rather than an event memory: it helps preserve identity, appearance, and scene state, but does not reliably support retrieval of specific past events. Finally, *(3) what attention mechanism underlies this effect?* Attention analyses under both fixed and changing prompts reveal that attention to middle KV is broadly distributed across the historical context, rather than collapsing onto a few isolated frames or prompt-transition boundaries.

Based on these findings, we further propose Temporal-Weighted Reservoir (TWR), a training-free KV-cache policy for autoregressive video generation. TWR builds on reservoir sampling, which maintains a fixed-size subset from an arbitrarily long stream by giving historical states a continued chance of being retained as new states arrive. This allows TWR to preserve broad coverage of middle KV over an unknown-length generation history without introducing deterministic temporal blind spots. Besides, we propose an extra time-weighted survival rule that biases retention toward earlier visual references. Extensive evaluations across a diverse set of autoregressive video backbones show that TWR consistently improves long-horizon visual consistency under matched KV-cache budgets. More importantly, TWR validates our analysis of middle KV in practice: preserving broad historical coverage while favoring earlier visual references leads to stronger long-horizon identity and appearance consistency. These results show that understanding the functional role of middle KV provides a useful basis for designing effective KV-cache policies for long-horizon autoregressive video generation.

In summary, our contributions are:

- We systematically investigate middle KV from the perspectives of its functional role and underlying attention mechanism, revealing that it serves as a visual reference rather than event memory and that attention to middle KV remains broadly distributed.
- We propose TWR, a novel training-free KV-cache policy that uses reservoir sampling to maintain broad coverage of middle history and a time-weighted survival rule to preferentially preserve earlier visual references over an unknown-length middle streaming history.
- Extensive experiments across diverse autoregressive video backbones show that TWR achieves consistently strong performance under matched KV-cache budgets. Beyond validating the effectiveness of TWR, these results further support our functional analysis of middle KV and demonstrate its practical value for principled KV-cache design.

2 Background and Related Work

Autoregressive video diffusion. Autoregressive video diffusion models generate frames or blocks causally and reuse past keys and values, making streaming and variable-length synthesis practical. CausVid distills a bidirectional video diffusion model into a few-step causal generator, while Self Forcing trains on the model's own rollouts to reduce exposure bias (Yin et al., 2025; Huang et al., 2025a). LongLive extends this paradigm through streaming long tuning and KV recaching, whereas Rolling Forcing uses rolling-window joint denoising with an initial anchor (Yang et al., 2026; Liu et al., 2026). During a long rollout, the model must condition on imperfect self-generated states under a bounded cache budget. Useful history can be evicted, while errors in retained history can propagate to later frames. Prior analysis describes these effects as history forgetting and temporal degradation, with degradation accumulating from per-step prediction errors (Wang et al., 2026).

KV-cache structures for long-video generation. Existing bounded-cache designs retain only recent history, combine an initial sink with a recent window, or use multi-timescale structures. Self Forcing uses a rolling cache of recent states, whereas LongLive and Rolling Forcing retain an initial sink together with a recent window (Huang et al., 2025a; Yang et al., 2026; Liu et al., 2026). Several methods explicitly organize retained history into early, intermediate, and recent regions. Deep

Forcing uses sink, candidate, and recent regions; Context Forcing uses sink, slow, and fast memory (Yi et al., 2026a; Chen et al., 2026b). Other hierarchical memory systems also combine intermediate history with local context (Wu et al., 2026). We use *sink*, *middle*, and *recent* as a common vocabulary for comparing these cache structures.

Historical-context management. Selection-based policies often use content signals to retain useful historical states. Deep Forcing keeps candidate KV tokens that participate in recent attention, whereas Context Forcing selects feature changes for slow memory (Yi et al., 2026a; Chen et al., 2026b). BIFE selects salient tokens within each block and retrieves sparse historical KV blocks by prompt similarity (Zhang et al., 2025). PackCache compacts historical states using condition anchoring, cross-frame decay, and spatially preserving position embeddings (Li et al., 2026). LongLive-RAG indexes generated latents as searchable history and retrieves top-k latents with a learned query embedding for each new block (Hu et al., 2026). WorldKV uses camera and action correspondence, while head-aware policies apply different cache rules across attention heads (Yi et al., 2026b; Chen et al., 2026a). Related LLM cache policies retain recent tokens together with persistent anchors or attention heavy hitters, and may select historical positions separately for different attention heads (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024). Other training-free policies retain pivotal tokens probabilistically or evict the least-attended state at each decoding step (Liu et al., 2023; Oren et al., 2024). In autoregressive video diffusion, Ca2-VDM shares conditional-frame caches across denoising steps, while Quant VideoGen reduces the numerical precision of retained KV states (Gao et al., 2025; Xi et al., 2026). Cache structure and retention mechanism are separate design choices; similar layouts with sink, middle, and recent regions can use different selection, compression, update, or retrieval policies.

Historical context in autoregressive video generation. Prior work provides relatively clear interpretations of the two endpoints of retained history. Initial states serve as stable global anchors, while recent states support local temporal continuity (Yang et al., 2026; Liu et al., 2026). Intermediate history is used in several ways. Deep Forcing treats highly attended intermediate states as extended anchors, whereas Context Forcing favors distinctive state transitions (Yi et al., 2026a; Chen et al., 2026b). These methods provide practical mechanisms for using intermediate history, but do not systematically investigate what information middle KV provides during generation. We systematically investigate the functional role and underlying attention mechanism of middle KV across diverse autoregressive video backbones.

3 The Fundamental Nature of Middle KV

Following the decomposition in the introduction, we divide historical video KV into three functional regions: sink, recent, and middle. The sink is the first frame, a clean historical anchor before autoregressive drift accumulates, while recent KV supports local continuity. Middle KV spans much of the long-range history, but its functional role has received limited systematic study.

We study middle KV through three questions. Does retaining it causally improve long-range visual consistency when sink and recent KV are fixed? Once its causal value is established, does middle KV act as matched event memory or as a continuous visual reference? Finally, does attention to middle KV concentrate on a small set of temporal locations or remain broadly distributed across middle history under fixed and changing prompts?

To study these questions with a precise and comparable readout, we use controlled test suites and focus on a representative form of long-horizon visual drift. Faces are visually important in generated videos and are particularly sensitive to accumulated drift, so preserving a subject's face has a direct impact on video quality. Face drift can also be measured more precisely than general visual drift: generic video metrics often mix changes in identity, motion, background, and detection, whereas face localization and ArcFace similarity (Deng et al., 2019) provide a direct and comparable measure of facial identity preservation. We therefore use ArcFace-based similarity as the main readout in our mechanism studies and construct the long-horizon test suites from single-person I2V cases in VBench-I2V (Huang et al., 2025b) with a clear target face. We use the no-switch and prompt-switch test suites for the long-horizon analyses and the sparse-event test suite for the event-retrieval probe. Appendix A.2 details the input construction, and Appendix B defines the metrics used in the main figures.

We assess the cross-backbone generality of these findings on four representative autoregressive video backbones: daVinci-MagiHuman (SII-GAIR et al., 2026), Wan2.1 and Wan2.2 from the Wan model family (Wan Team et al., 2025), and the HY-World 1.5 (Tencent Hunyuan Foundation Model Team, 2025; Sun et al., 2025). Appendix A.1 describes the corresponding platforms, execution paths, and backbone-local KV budget definitions. Across all models, cache interventions are executed at the granularity of latent-frame KV slots.

3.1 Middle KV Has Causal Value for Long-Range Consistency

Controlled middle KV visibility comparison. We first ask whether middle KV has a causal effect once the sink and recent window are fixed. Under matched prompts and generation seeds, we compare a no-middle setting that hides the middle region with a with-middle setting that restores it. We evaluate both settings over the first 60 seconds on the no-switch and prompt-switch test suites.

As shown in Figure 1, restoring middle-history visibility increases identity similarity across all four backbones over the first 60 seconds. The 95% paired-bootstrap confidence interval for the 0–60s mean difference remains above zero for every backbone on both test suites. Middle KV therefore has a clear causal effect on long-range visual consistency. This result shows directly that long-video cache retention needs access to the middle region, not only the sink and recent KV.

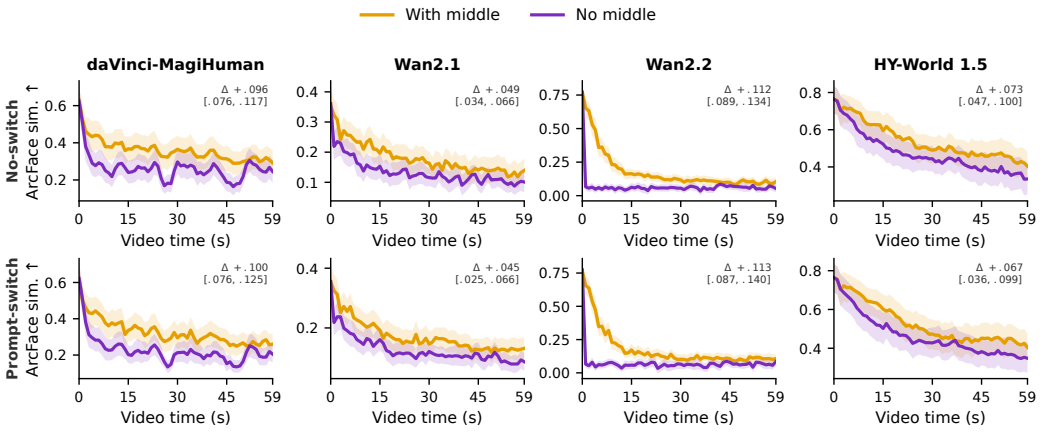

Figure 1: Time-resolved ArcFace similarity over 60 seconds on the no-switch (top) and prompt-switch (bottom) test suites. Orange restores middle KV; purple hides it with the recent window fixed. Curves average valid target-face samples from the complete set of 39 cases. Shading and insets show pointwise and 0–60s paired-bootstrap 95% confidence intervals, respectively. The retention seed is fixed to zero.

3.2 Middle KV Provides Visual Reference Rather Than Event Memory

After establishing that middle KV matters, we ask what information it provides. One possibility is event memory: the model might retrieve a few matched past events based on attention scores or semantic similarity. Another possibility is visual reference: the model may use middle KV as a continuous record of identity, appearance, and scene state.

Sparse-event probes show no stable matched-event retrieval from middle KV. We test the event-memory view on the sparse-event test suite. Each case prompts two occurrences of the same action, such as using a phone, with the two event windows placed according to one of four schedules: A, B, C, or D. If middle KV works as event memory, the query during the second occurrence should produce a schedule-invariant attention peak on the first occurrence. We evaluate this probe on daVinci-MagiHuman, Wan2.1, and Wan2.2. HY-World 1.5 is excluded because its long-video generations do not follow the prompted sparse actions reliably enough to support this event schedule. Figure 2 shows that the second event does not reliably induce an attention peak on the earlier matching event. Attention follows the model's native temporal layout more than semantic correspondence, so mid-

dle KV does not behave like a sparse event-addressable lookup table. Appendix C reports the same analysis for a cup event.

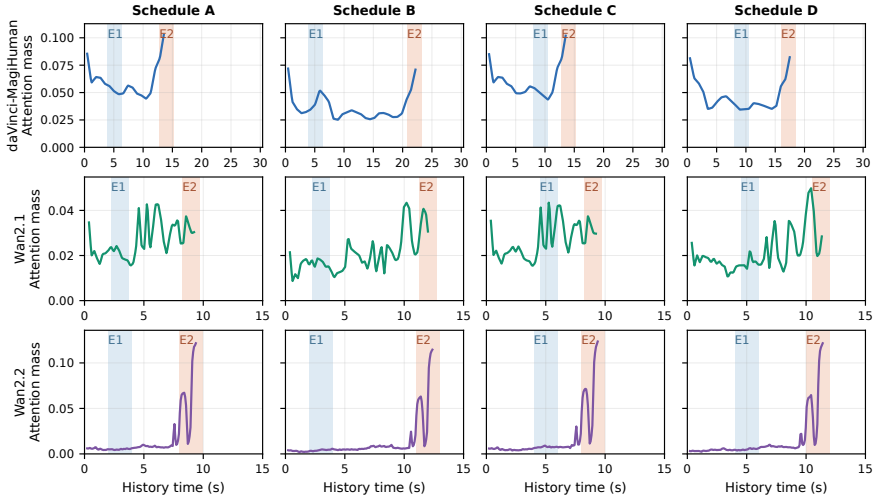

Figure 2: Sparse phone-event readout on the sparse-event test suite. Rows show backbones and columns show the four schedules. Shading marks event windows; curves report normalized historical attention for the second-event query. The absence of a schedule-invariant peak on the earlier match indicates that historical KV is not primarily retrieved as matched event memory.

Controlled corruption experiments show that middle KV serves as a visual reference for subsequent generation. We perturb retained middle KV and observe later frames. Face-swap replay writes wrong-identity content, while horizontal-strip replay permutes latent bands. The sink, recent window, current block, and unselected states remain clean, and larger doses add nested corrupted states. We report the response at each corruption dose under matched generation settings. Both perturbations propagate to later generated frames (Figure 3), confirming that middle KV serves as visual context for subsequent generation.

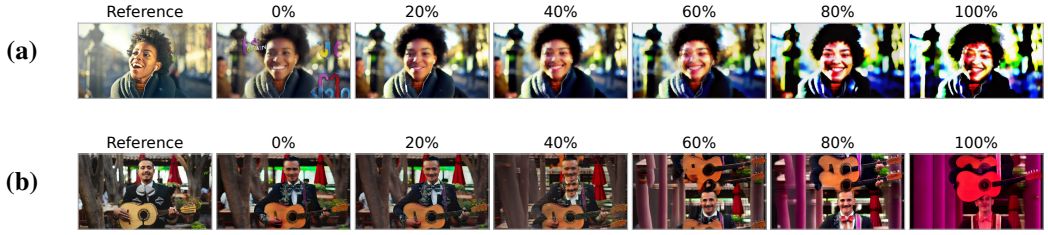

Figure 3: History-contamination replay at 48s. Each row shows the reference and matched frames with 0–100% contamination. (a) Face-swap contamination changes identity. (b) Horizontal-strip contamination displaces latent bands. Both changes propagate to later generated frames, showing that middle KV serves as a visual reference.

Appendix Figures 12 and 13 confirm that similarity falls as corruption increases. Together with the sparse-event probe, this identifies middle KV as a continuous visual reference for identity, appearance, and scene structure rather than a matched event index. Because this self-generated state can carry errors forward, retention should preserve broad coverage while favoring earlier, cleaner anchors.

3.3 Middle KV Attention Is Broadly Distributed, Not Sparse

Historical attention spans broad temporal regions. We finally ask whether attention to middle KV concentrates on a few temporal locations or remains broadly distributed. Using the no-switch and

prompt-switch test suites, we measure historical attention over absolute time and average the readout over model components and test cases. The prompt-switch readout uses the shared-schedule control described in Appendix A.3, whereas the policy comparisons use case-specific schedules. We use backbone-specific rollout durations because the backbones differ in how long they maintain stable visual quality. Figure 4 shows broad temporal bands under fixed prompts. Under changing prompts, attention also remains distributed rather than concentrating at prompt-switch boundaries. This pattern supports treating middle KV as long-range visual context and motivates retention policies that preserve broad temporal coverage. Similar temporal patterns also appear at the layer, attention-head, and individual-case levels, as shown in Appendix D.

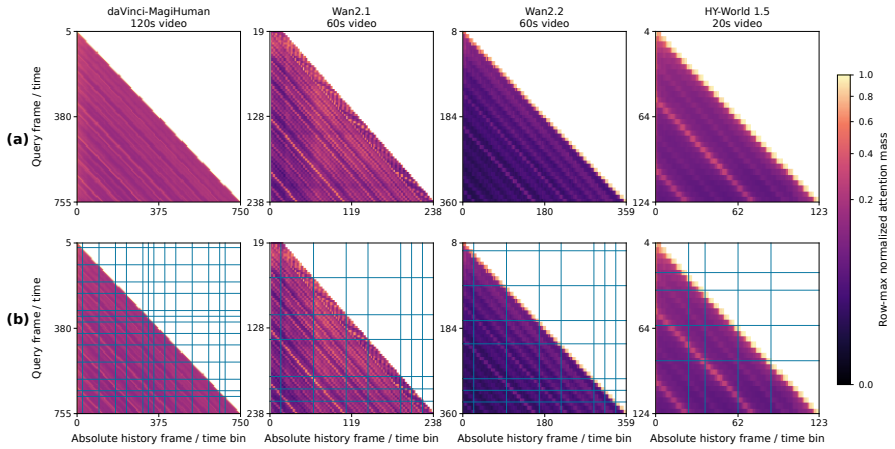

Figure 4: Attention to historical KV on the (a) no-switch and (b) prompt-switch test suites; cyan lines mark prompt-switch boundaries. Each heatmap reports attention from later query frames to earlier historical KV slots, averaged over model components and all cases in the corresponding test suite. Attention spans broad temporal regions in both settings rather than concentrating at the switches.

4 Temporal-Weighted Reservoir

The earlier analyses show that cache retention needs both broad coverage of middle history and a preference for earlier states. Because the final video length is unknown during generation, reservoir sampling processes the history online with fixed memory while continuing to cover non-recent states. Temporal- Weighted Reservoir, abbreviated TWR, keeps this coverage and favors earlier candidates without training. Figure 5 shows the update process.

4.1 Reservoir Coverage

TWR divides a fixed latent-frame KV budget B into S sink slots, a recent window of size R, and a middle reservoir with $M = B - S - R$ slots. The sink keeps the initial visual anchor, the recent window supports local continuity, and the middle reservoir retains non-recent states. All formal TWR and Uniform Reservoir settings use $S = 1$ and $R = 5$, with a backbone-specific total budget B. New states enter the recent window in chronological order. When it grows beyond R, the oldest state becomes a middle candidate; if a model call adds several frames, candidates are processed chronologically with one random draw each.

Uniform Reservoir applies Algorithm R to this candidate stream (Vitter, 1985). The first M candidates fill the middle reservoir. For candidate $n > M$, the algorithm draws j uniformly from $\{0, \ldots, n-1\}$, replaces middle slot j if $j < M$, and otherwise discards the candidate. Thus, after n candidates, every eligible state has marginal retention probability M/n. Selected middle states are sorted by frame index and written back as

$$[\text{sink} \mid \text{sorted middle states} \mid \text{recent}].$$

This equal-weight baseline isolates the value of broad historical coverage under the same cache layout and candidate stream.

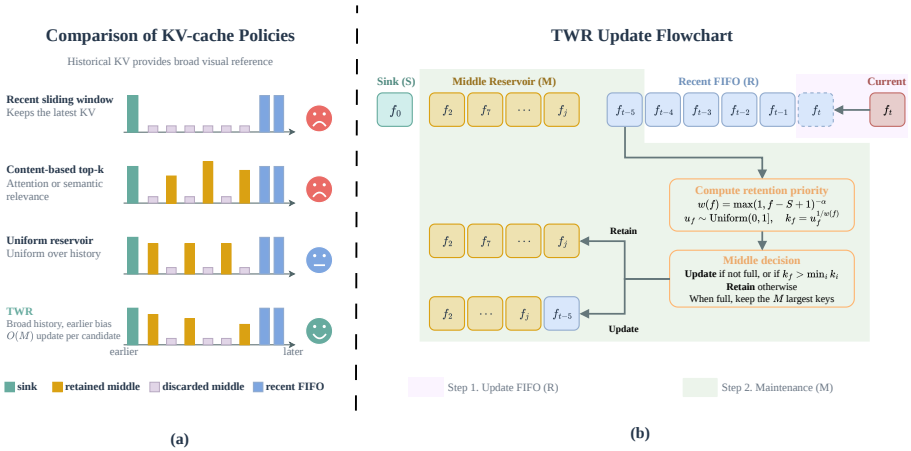

Figure 5: **Temporal-Weighted Reservoir.** (a) Comparison with common KV-cache policies. (b) TWR updates the middle reservoir using temporally weighted random keys.

4.2 Early-Time Survival Weighting

Model-generated cached states continue to affect later frames. Section 3.2 shows that corrupted middle states can propagate identity and structural drift. Earlier states are generated closer to the observed input and pass through fewer autoregressive steps, reducing their exposure to accumulated generation errors. TWR therefore favors their survival while continuing to sample from the full non-recent history.

Let f be the zero-indexed absolute latent-frame index; sink frames do not enter the candidate stream. TWR assigns each eligible middle state the weight

$$w(f) = \max(1, f - S + 1)^{-\alpha}.$$

It then draws $u_f \sim \text{Uniform}(0, 1]$ and computes the Efraimidis–Spirakis key (Efraimidis & Spirakis, 2006)

$$k_f = u_f^{1/w(f)}.$$

For $\alpha > 0$, earlier candidates receive larger weights and are more likely to receive large keys.

The first M candidates fill the reservoir. Once it is full, let $\tau = \min_{i \in \mathcal{M}} k_i$ be its smallest key. A new candidate is admitted only when $k_f > \tau$, in which case it replaces the state with key τ. After any prefix of the candidate stream, the middle reservoir contains the M largest keys seen so far; an admitted state may still be replaced by a later candidate. At the current threshold, candidate f is admitted with probability $1 - \tau^{w(f)}$.

For numerical stability, the implementation compares the equivalent log-domain score

$$\log w(f) - \log(-\log u_f).$$

A linear scan over the resident scores gives $O(M)$ time per candidate and $O(M)$ storage. Under a fixed cache budget, this update cost is independent of the number of previously seen candidates. The update uses only the frame index, one random draw, and the resident scores, without content-dependent scoring. Formal TWR uses $\alpha = 1$, which favors earlier candidates while leaving every eligible candidate a nonzero admission probability. TWR and Uniform Reservoir share the cache layout, candidate stream, retention seeds, and chronological writeback, so their comparison isolates early-time weighting. The four backbone integrations use the same logical selection rule.

5 Policy Evaluation

We evaluate retention policies on the 39-case no-switch test suite used for the mechanism probes. Within each backbone, comparisons match the prompt, generation seed, duration, generation set-

tings, and active KV budget. This budget counts the resident latent-frame KV slots used by attention. All policies retain KV at the granularity of latent-frame slots and reserve one sink slot. Policies with an explicit middle region use a five-slot recent window. The sink+fifo baseline instead assigns all remaining slots to its recent FIFO queue. Uniform Reservoir uses Algorithm R, while TWR uses weighted reservoir keys with $\alpha = 1$. Their contrast isolates early-time survival weighting under the same cache geometry and candidate stream. Main-table results for Uniform Reservoir and TWR average six fixed retention seeds, numbered 0 through 5, within each case. The deterministic baselines are evaluated once. Budget curves use retention seed 0 for Uniform Reservoir and TWR. We use backbone-specific evaluation horizons to keep the long-form comparison meaningful, evaluating daVinci-MagiHuman, Wan2.1, and Wan2.2 up to 240 seconds, and HY-World 1.5 up to 120 seconds. ID drift follows Appendix B; lower is better.

Table 1: Long-horizon comparison on the no-switch test suite at 240s for daVinci-MagiHuman, Wan2.1, and Wan2.2, and 120s for HY-World 1.5. Scores use a 0–100 scale. Arrows indicate the preferred direction; Dynamic is descriptive. Uniform Reservoir is the equal-weight counterpart of TWR. Uniform Reservoir and TWR aggregate six retention seeds. All methods match the active KV budget. LongLive-RAG's external searchable KV archive is not included in this budget.

Backbone	Strategy	FaceSim↑	ID drift↓	Face det.↑	DINO↑	Scene drift↓	Image Q↑	Q drift↓	Dynamic	Motion↑
daVinci-MagiHuman	sink+fifo	18.5	19.6	99.7	76.3	5.4	**70.0**	3.6	20.6	99.4
	Deep Forcing	18.8	23.1	99.6	76.4	5.3	69.9	3.9	21.7	99.4
	Uniform reservoir	34.9	12.5	**100.0**	83.8	2.0	69.8	2.4	5.9	**99.6**
	TWR	**42.1**	**4.4**	**100.0**	**85.7**	**0.8**	69.7	**1.5**	5.4	**99.6**
Wan2.1	sink+fifo	11.9	22.2	70.8	44.3	31.4	47.3	30.0	14.7	96.2
	Deep Forcing	13.7	22.8	38.7	31.8	36.0	47.6	29.2	26.1	93.7
	Uniform reservoir	22.2	7.5	**100.0**	75.0	9.8	69.3	5.8	4.5	**99.3**
	LongLive-RAG	14.6	23.5	82.3	54.7	26.8	57.1	28.9	11.3	97.0
	TWR	**26.8**	**2.9**	**100.0**	**80.5**	**2.5**	**71.7**	**2.6**	6.8	99.2
Wan2.2	sink+fifo	19.0	44.2	69.3	48.0	29.1	49.6	19.8	28.4	98.3
	Deep Forcing	28.2	38.4	25.2	26.2	32.6	32.2	33.7	32.6	96.7
	Uniform reservoir	30.1	33.3	97.9	76.2	8.9	57.5	8.9	15.0	99.1
	TWR	**42.9**	**24.1**	**99.9**	**84.6**	**3.6**	**60.6**	**3.7**	14.2	**99.4**
HY-World 1.5	sink+fifo	44.3	64.6	72.6	57.7	38.0	65.7	11.5	10.0	99.0
	Deep Forcing	48.2	66.6	64.5	54.7	37.4	65.1	15.7	8.7	98.8
	Uniform reservoir	53.4	39.9	87.3	77.5	**17.0**	**69.8**	5.6	6.1	99.2
	TWR	**56.0**	**36.3**	**91.0**	**80.2**	**17.0**	69.7	**5.4**	5.4	**99.4**

5.1 Main Results

Table 1 compares retention under matched active KV budgets across four backbones. In table-row order, sink+fifo preserves the fixed sink and assigns every remaining KV slot to a FIFO queue of the most recent states. Deep Forcing applies score-based Participative Compression, while Uniform reservoir provides unweighted coverage over eligible history. We report LongLive-RAG (Hu et al., 2026) only on Wan2.1 because our validated implementation uses its released retrieval encoder in the Wan2.1 latent space. The base generator remains frozen. Applying it to other backbones would require separate encoder adaptation and validation. TWR adds early-time survival weighting to the same reservoir geometry.

At the longest no-switch horizon, TWR achieves the highest FaceSim and lowest ID drift on all four backbones. Compared with Uniform Reservoir, it improves DINO similarity on every backbone, reduces or matches scene drift, lowers quality drift, and maintains or improves face detection. Image quality and motion smoothness remain comparable. On Wan2.1, TWR also outperforms LongLive-RAG on all metrics with an indicated direction. The prompt-switch results show the same broad pattern, with TWR leading in FaceSim, ID drift, DINO similarity, scene drift, and quality drift at the longest horizon on all four backbones. Full results at 30 seconds, 60 seconds, and the longest horizon are reported in Appendix E. Figure 6 shows one Wan2.1 60s case.

5.2 Budget Scaling

Figure 7 tests whether a larger recent window alone is sufficient for long-horizon identity stability. It compares the four retention rules on matched budget grids for the no-switch test suite. The figure reports ID drift at 240 seconds for daVinci-MagiHuman, Wan2.1, and Wan2.2, and at 120 seconds for HY-World 1.5. Lower is better. Appendix B defines the metric and missing-sample handling.



Figure 6: Wan2.1 60-second generations from the no-switch test suite under an active KV budget of $B = 45$. Rows show retention policies; columns sample 0, 15, 30, 45, and 60 seconds.

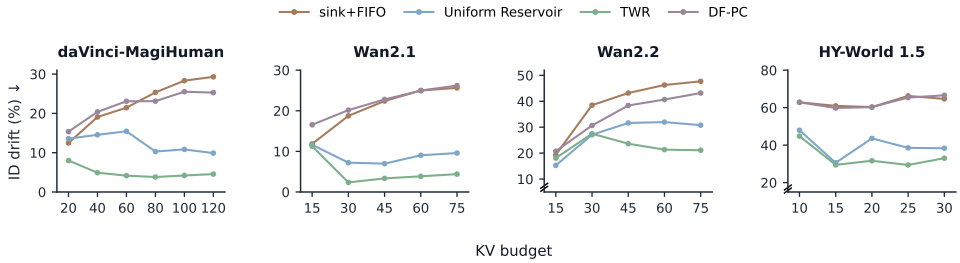

Figure 7: Budget scaling on the no-switch test suite at each backbone's longest horizon. TWR keeps ID drift lower than recent-only expansion and score-based compression. Metric details are in Appendix B.

Across all evaluated budgets and backbones, TWR has lower ID drift than sink+fifo and Deep Forcing. Uniform Reservoir often improves over recent-only retention, showing the value of broad historical coverage. TWR further improves over Uniform Reservoir at most budget points, showing the additional benefit of favoring earlier visual references. These results support the design introduced in Section 1. Broad coverage of middle history and preferential retention of earlier visual references are both important for long-horizon consistency.

6 Conclusion

This paper systematically investigates the functional nature of middle KV in autoregressive video generation across diverse backbones. Our analysis shows that retaining middle KV has a causal effect on long-range visual consistency. Middle KV serves primarily as a visual reference rather than event memory, helping preserve identity, appearance, and scene state without reliably supporting the retrieval of specific past events. Attention to middle KV is broadly distributed across the historical context rather than concentrating on a few isolated frames or prompt-transition boundaries.

Guided by these findings, we introduced Temporal-Weighted Reservoir (TWR), a training-free KV-cache policy that uses reservoir sampling to maintain broad coverage of middle KV over an unknown-length generation history without deterministic temporal blind spots. A time-weighted survival rule further favors earlier visual references. Across diverse autoregressive video backbones, TWR consistently improves long-horizon visual consistency under matched KV-cache budgets. Together, these results show that understanding the functional role of middle KV provides a useful basis for principled KV-cache design in long-horizon autoregressive video generation.

Ethics statement. This work uses existing model checkpoints and benchmark inputs and involves no human participants or new personal data. Face embeddings are used only to evaluate visual consistency, and the controlled face-swap intervention studies how identity information propagates through cached states. Video generation can be misused to create deceptive or harmful content. We support use of this work in accordance with applicable licenses, privacy requirements, and the ICLR Code of Ethics.

Reproducibility statement. We provide the details needed to reproduce our results in the paper and appendix. Appendices A and A.2 document the checkpoints, generation settings, cache budgets, policy configurations, frozen inputs, prompts, and evaluation durations. Appendix B specifies the retention seeds, sampling rules, metrics, aggregation, missing-value handling, and statistical analysis. We plan to release the code, configurations, prompts, manifests, and evaluation scripts upon publication.

AI use statement. We used generative AI tools to aid and polish the writing of this paper. The authors reviewed and edited all AI-assisted text and verified the final claims and citations. The authors take full responsibility for the content of the paper.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Jiayu Chen, Junbei Tang, Wenbiao Zhao, Maoliang Li, Jiayi Luo, Zihao Zheng, Jiawei Yang, Guojie Luo, and Xiang Chen. Pyramid forcing: Head-aware pyramid kv cache policy for high-quality long video generation, 2026a. URL https://arxiv.org/abs/2605.13111.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2602.06028.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, Yuyang Zhao, Huizi Mao, Ying-Cong Chen, Enze Xie, Xiaojuan Qi, and Song Han. Longlive-2.0: An NVFP4 parallel infrastructure for long video generation, 2026c. URL https://arxiv.org/abs/2605.18739.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Pavlos S. Efraimidis and Paul G. Spirakis. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181–185, 2006. doi: 10.1016/j.ipl.2005.11.003.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.00373. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Fang_Perceptual_Quality_Assessment_of_Smartphone_Photography_CVPR_2020_paper.html.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-VDM: Efficient autoregressive video diffusion model with causal generation and cache sharing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267

of *Proceedings of Machine Learning Research*, pp. 18550–18565. PMLR, 2025. URL https://proceedings.mlr.press/v267/gao25m.html.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. Longlive-rag: A general retrieval-augmented framework for long video generation, 2026. URL https://arxiv.org/abs/2606.02553.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

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*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Kunyang Li, Mubarak Shah, and Yuzhang Shang. Packcache: A training-free acceleration method for unified autoregressive video generation via compact kv-cache, 2026. URL https://arxiv.org/abs/2601.04359.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0722. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Kunhao Liu et al. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, 2026. URL https://kunhao-liu.github.io/Rolling_Forcing_Webpage/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2279.

URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a452a7c6c463e4ae8fbdc614c6e983e6-Abstract-Conference.html.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18724–18741. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1043. URL https://aclanthology.org/2024.emnlp-main.1043/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

SII-GAIR, Sand.ai, Ethan Chern, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model, 2026. URL https://arxiv.org/abs/2603.21986.

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

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419, 2020. doi: 10.1007/978-3-030-58536-5_24.

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Jeffrey S. Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985. doi: 10.1145/3147.3165.

Wan Team et al. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Jing Wang, Fengzhuo Zhang, Xiaoli Li, Vincent Y. F. Tan, Tianyu Pang, Chao Du, Aixin Sun, and Zhuoran Yang. Error analyses of auto-regressive video diffusion models. *Journal of Machine Learning Research*, 27(135):1–51, 2026. URL https://www.jmlr.org/beta/papers/v27/25-3128.html.

Mingqiang Wu, Weilun Feng, Zhefeng Zhang, Haotong Qin, Yuqi Li, Guoxin Fan, Xiaokun Liu, Zhulin An, Libo Huang, Yongjun Xu, and Chuanguang Yang. Echo-forcing: A scene memory framework for interactive long video generation, 2026. URL https://arxiv.org/abs/2605.16003.

Haocheng Xi, Shuo Yang, Yilong Zhao, Muyang Li, Han Cai, Xingyang Li, Yujun Lin, Zhuoyang Zhang, Jintao Zhang, Xiuyu Li, Zhiying Xu, Jun Wu, Chenfeng Xu, Ion Stoica, Song Han, and Kurt Keutzer. Quant videogen: Auto-regressive long video generation via 2-bit KV-cache quantization, 2026. URL https://arxiv.org/abs/2602.02958.

Guangxuan Xiao et al. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

Shuai Yang et al. Longlive: Real-time interactive long video generation. In *International Conference on Learning Representations*, 2026. URL https://nvlabs.github.io/LongLive/.

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. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2512.05081.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. Worldkv: Efficient world memory with world retrieval and compression, 2026b. URL https://arxiv.org/abs/2605.22718.

Tianwei Yin et al. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://causvid.github.io/.

Zeyu Zhang, Jinyuan Mao, Shuning Chang, Yuanyu He, Yizeng Han, Jiasheng Tang, Fan Wang, and Bohan Zhuang. BIFE: Better interaction, fewer errors for minute-long video generation, 2025. URL https://arxiv.org/abs/2511.22973v2. Version 2, revised June 2026.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract.html.

Appendix Contents

- Appendix A: Evaluation Setup
- Appendix B: Evaluation Protocol
- Appendix C: Sparse Event Readout
- Appendix D: Additional Historical Attention Readouts
- Appendix E: Extended Policy Results
- Appendix F: Controlled History Interventions

A Evaluation Setup

A.1 Models and Cache Budgets

We evaluate cache policies on four frozen autoregressive video generation systems. Each system receives an initial image and a text prompt. The policy changes only which historical KV slots remain available during generation.

daVinci-MagiHuman. We use the official MagiHuman model configuration and its frozen same-noise-level streaming checkpoint (SII-GAIR et al., 2026). The model is a 40-layer causal video diffusion transformer. Generation uses four denoising steps, a streaming VAE, and 25 FPS output.

Wan2.1. We use the Wan2.1-T2V-1.3B network with the frozen Self-Forcing DMD checkpoint and run it in image-to-video mode (Wan Team et al., 2025; Huang et al., 2025a; Wu et al., 2026). Generation uses four denoising steps, the streaming VAE path, and 16 FPS output.

Wan2.2. We use the Wan2.2-TI2V-5B network with the frozen LongLive2 checkpoint in image-to-video mode (Wan Team et al., 2025; Chen et al., 2026c). Generation uses four sampling steps. The initial image is encoded with the native Wan VAE, and the generated frames are decoded with MG-LightVAE v2.

HY-World 1.5. We use the HunyuanVideo-1.5 base checkpoint together with the frozen HY-WorldPlay action-distilled autoregressive checkpoint (Tencent Hunyuan Foundation Model Team, 2025; Sun et al., 2025). This combined execution is called HY-World 1.5 throughout the paper. Generation follows the frozen 24 FPS runtime settings.

The budget B is the maximum number of resident latent-frame KV slots, including the fixed sink. The current generation block remains available to the model. A cache policy selects the historical

Table 2: Backbone-specific budget and cache settings.

Backbone	Budget parameter	Budget-scaling grid	Generation seed
daVinci-MagiHuman	window_frames	$\{20, 40, 60, 80, 100, 120\}$	42
Wan2.1	local_attn_size	$\{15, 30, 45, 60, 75\}$	0
Wan2.2	local_attn_size	$\{15, 30, 45, 60, 75\}$	0
HY-World 1.5	memory_frames	$\{10, 15, 20, 25, 30\}$	42
Backbone	**One budget unit**		**Generation block**
daVinci-MagiHuman	One latent-frame KV slot		0.8s
Wan2.1	One slot with 1560 spatial tokens		Three slots, 0.75s
Wan2.2	One slot with 880 spatial tokens		Eight slots, 1.333s
HY-World 1.5	One latent-frame KV slot		Native memory-frame block

slots stored within the remaining capacity. Table 2 separates the parameter used for B from the meaning of one budget unit and the generation block size. Equal values of B across backbones do not imply equal token counts, history duration, or computation. We therefore compare policies only within a backbone under a matched budget.

The policy comparisons in Section 5 and Appendix E use one fixed budget per backbone: $B = 60$ for daVinci-MagiHuman, $B = 45$ for Wan2.1 and Wan2.2, and $B = 30$ for HY-World 1.5. The full grids in Table 2 are used for the budget-scaling analysis.

Within each backbone and input suite, all compared policies use the same initial frame, prompt sequence, generation seed, generation settings, evaluation endpoints, and total cache budget. The backbone-specific generation seed stays fixed across policies and is separate from the retention seeds used by stochastic cache policies.

A.2 No-Switch Test Suite

The no-switch test suite contains 39 frozen image-to-video cases derived from the VBench-I2V single-human pool (Huang et al., 2025b). The cases use identifiers `case00` through `case38`. Each case fixes the first frame, the underlying English event, the case identifier, and the backbone-specific prompt adapter before generation. The selected first frames contain one clearly identifiable person. All backbones and cache policies use the same 39 underlying cases.

The prompt remains fixed throughout each generated video. The four backbones receive the same underlying event through their frozen native prompt formats. Generation lasts 240.8s for daVinci-MagiHuman, 240.75s for Wan2.1, 241.333s for Wan2.2, and 121s for HY-World 1.5. The evaluated prefixes are 30s, 60s, and 240s for the first three backbones, and 30s, 60s, and 120s for HY-World 1.5.

A.3 Prompt-Switch Test Suite

The prompt-switch test suite reuses the same 39 frozen first frames. Each case has a scene-compatible sequence of actions. Segment lengths are sampled reproducibly from 4s, 8s, and 12s, and the complete timeline is frozen before generation. For a given case, all cache policies receive the same action sequence and switch times. The nominal action identities and switch times are also shared across backbones, while the wording follows each backbone's native prompt format.

Case-specific schedules. The policy comparison uses a different frozen timing and action sequence for each case. This prevents one generic sequence from imposing incompatible motions on all scenes. Native generation boundaries can shift a nominal switch slightly. In Wan2.1, which advances in 0.75s blocks, the largest shift is 0.25s. The final native generation step repeats the last prompt and does not introduce a new action.

Shared schedule. The aggregate temporal readout uses a separate control in which all 39 cases share the same segment boundaries. Action text remains specific to each scene. This aligned sched-

ule makes prompt boundaries comparable across cases. It is used for the temporal readout analysis, while the policy tables use the case-specific schedules above.

A.4 Sparse-Event Test Suite

The sparse-event test suite uses 15 cases from the frozen 39-case pool. These cases were selected because their initial frames can support the prescribed foreground event, making the event easier to observe across schedules.

Each video contains two occurrences of the same foreground event with an idle interval between them. Four frozen schedules, denoted A through D, move the event windows while keeping the event meaning fixed within each backbone and case. The same case and schedule use matched prompts and generation settings when historical readout is compared.

B Evaluation Protocol

B.1 Sampling and Case Aggregation

The policy tables report nine metrics. FaceSim, DINO similarity, Face detection coverage, Scene drift, and Image quality use a planned 1 FPS sampling grid at $0.5, 1.5, \ldots, D - 0.5$ seconds for a prefix of duration D. Timestamp lookup uses the nearest encoded frame. Dynamic degree and Motion smoothness use the frozen VBench sampling and normalization code (Huang et al., 2024).

A metric is first computed for each case and then averaged with equal case weight. For method a at duration D,

$$\bar{m}_{a,D} = \frac{1}{39} \sum_{c=1}^{39} m(V_{a,c,D}).$$

The reported policy tables use the complete set of 39 cases for every method in the same backbone and duration. Frames are never pooled across cases. This prevents frame rate, video length, or detection frequency from changing the weight of a case.

The final tables contain a finite case-level value for every reported method, metric, and case. Outside the detection-specific rules below, missing or nonfinite case-level values are not imputed. An incomplete evaluation must be repaired and pass the 39-case completeness check before aggregation.

All paper-facing table values are multiplied by 100. The archived metric files retain their native scales. Higher values are better for FaceSim, DINO similarity, Face detection coverage, Motion smoothness, and Image quality. Lower values are better for ID drift, Scene drift, and Quality drift. Dynamic degree is descriptive and has no preferred direction.

B.2 Stochastic Cache Policies

Uniform Reservoir and TWR use stochastic retention. Both policies use the six fixed retention seeds $\{0, 1, 2, 3, 4, 5\}$ for every backbone, input, budget, and case. The generation seed in Table 2 stays fixed, so the retention seed changes only cache admission and replacement decisions.

Let $m_{c,r}$ be one case-level metric for retention seed r. We first compute

$$\bar{m}_c = \frac{1}{6} \sum_{r=0}^{5} m_{c,r},$$

and then average $\bar{m}_c$ over the 39 cases. The six runs are repeated retention samples, not independent cases. The original seed-level results are retained. Deterministic policies are executed once per case and are not duplicated across retention seeds.

B.3 Identity Metrics

FaceSim. At each planned 1 FPS timestamp, we detect and align the target face and extract an ArcFace R100 embedding (Deng et al., 2019). The reference and generated embeddings are L2

normalized. If $\mathcal{F}_{c,D}$ is the set of samples with a valid target-face match, then

$$\mathrm{FaceSim}(V_{c,D}) = \frac{1}{|\mathcal{F}_{c,D}|} \sum_{t \in \mathcal{F}_{c,D}} \left\langle \hat{e}_c^{\mathrm{ref}}, \hat{e}_{c,t} \right\rangle .$$

Detection failures, faces that are too small, and samples without a reliable target association are excluded from this similarity mean and counted by Face detection coverage. Higher is better.

ID drift. ID drift is the signed decrease in FaceSim from the beginning to the end of a video. Let $\mathcal{T}_D = (t_0, \ldots, t_{N-1})$ be the planned 1 FPS sampling times and let $n_D = \lceil 0.15N \rceil$. The early window contains the first n_D samples and the late window contains the last n_D samples. The first planned video sample is included. For each planned sample, define

$$\tilde{s}_t = \begin{cases} s_t, & \text{the target face is reliably matched and } s_t \text{ is finite,} \\ 0, & \text{otherwise.} \end{cases}$$

The table reports the signed difference

$$\mathrm{IDDrift}(V_{c,D}) = \mathrm{mean}_{t \in W_{\mathrm{early}}} \tilde{s}_t - \mathrm{mean}_{t \in W_{\mathrm{late}}} \tilde{s}_t.$$

Lower is better. A positive value means that identity similarity is lower near the end of the video. Missing or unreliable target faces contribute zero to this metric instead of being removed from the two fixed windows.

Face detection coverage. Face detection coverage uses the same planned samples and target association as FaceSim,

$$\mathrm{FaceCov}(V_{c,D}) = \frac{|\mathcal{F}_{c,D}|}{|\mathcal{T}_D|}.$$

Higher is better. Reporting coverage with FaceSim separates identity similarity on valid target faces from the frequency with which the target face remains visible and identifiable.

B.4 Content, Motion, and Quality Metrics

DINO similarity. A frozen subject phrase is used with Grounding DINO and SAM to isolate the reference subject and the generated subject (Liu et al., 2024; Kirillov et al., 2023). DINO ViT-B/16 features are L2 normalized (Caron et al., 2021). For the valid subject samples $\mathcal{S}_{c,D}$,

$$\mathrm{DINOSim}(V_{c,D}) = \frac{1}{|\mathcal{S}_{c,D}|} \sum_{t \in \mathcal{S}_{c,D}} \left\langle \hat{d}_c^{\mathrm{ref}}, \hat{d}_{c,t} \right\rangle .$$

Failed subject detections are excluded from the similarity mean and retained in the subject-validity diagnostics. Higher is better.

Dynamic degree. We use the official VBench Dynamic Degree evaluator (Huang et al., 2024). It uses RAFT flow (Teed & Deng, 2020) on sampled frame pairs and assigns each video a binary dynamic label. The reported value is the percentage of the 39 cases classified as dynamic, not mean optical-flow magnitude. It is descriptive; neither direction is universally preferable.

Motion smoothness. We use the official VBench Motion Smoothness evaluator. It predicts intermediate frames with the frozen interpolation model and compares them with the observed frames. If MAE is the mean absolute interpolation error, then

$$\mathrm{MotionSmoothness}(V) = \frac{255 - \mathrm{mean}(\mathrm{MAE})}{255}.$$

Higher is better. Dynamic degree and Motion smoothness are reported together so that low motion and unstable motion remain distinguishable.

Scene drift. A frozen CLIP image encoder maps the first and last 5s windows to normalized features (Radford et al., 2021). Each window contains five samples taken at 1 FPS. We average within each window, normalize the two means, and compute

$$\mathrm{SceneDrift}(V_{c,D}) = 1 - \langle \bar{c}_{\mathrm{early}}, \bar{c}_{\mathrm{late}} \rangle .$$

Lower is better. This is the paper's long-range window comparison and is reported as Scene drift rather than VBench Background Consistency.

Image quality. We evaluate frames on the planned 1 FPS sampling grid with the MUSIQ model trained on SPAQ, using the frozen VBench-compatible resize and normalization (Ke et al., 2021; Fang et al., 2020). If q_t is the MUSIQ score divided by 100, then

$$\mathrm{ImageQuality}(V_{c,D}) = \frac{1}{|\mathcal{T}_D|} \sum_{t \in \mathcal{T}_D} q_t.$$

Higher is better.

Quality drift. Using the same MUSIQ scores, we compare the first and last 5s windows,

$$\mathrm{QualityDrift}(V_{c,D}) = \left| \mathrm{mean}_{t \in W_{\mathrm{early}}} q_t - \mathrm{mean}_{t \in W_{\mathrm{late}}} q_t \right|.$$

Lower is better. The archived results also retain the signed early-minus-late quality difference, while the table reports the absolute change.

B.5 Statistical Reporting

Case-level paired bootstrap confidence intervals use cases as the resampling unit and preserve method pairing within every resample. For Uniform Reservoir and TWR, the six retention seeds for a case remain grouped during resampling. The result files retain the effective case count, the case-level values, the seed-level values for stochastic policies, and the seed-specific macro mean and standard deviation.

C Sparse Event Readout

The sparse event experiment tests whether the model reads an earlier event when the same event appears again. The idle interval separates the two events and reduces overlap between their local temporal contexts. The main text reports the phone event. Figure 8 reports the same analysis for a cup event.

We measure the attention assigned to the first event during the second event. We compare this value with the readout measured during nearby idle periods.

Historical readout changes with event placement and backbone-specific temporal structure. The repeated event does not produce a schedule-invariant peak at the earlier matching event.

D Additional Historical Attention Readouts

Additional layer- and head-specific readouts and case-level results under the no-switch and prompt-switch settings are shown below. They exhibit similar broad temporal patterns and support the aggregate analysis in the main text.

E Extended Policy Results

E.1 Results on the No-Switch Test Suite

Evaluation on the no-switch test suite measures long-horizon generation with a fixed prompt. It tests how cache retention affects identity, content, motion, and visual quality as generation continues.

TWR provides the strongest long-horizon identity preservation across the four backbones in the sealed comparison. At the longest evaluated duration for each backbone, it achieves the highest FaceSim and the lowest ID drift among the compared policies.

The accompanying motion and quality results show the behavior of each policy beyond identity preservation. Face detection coverage is reported with every identity result.

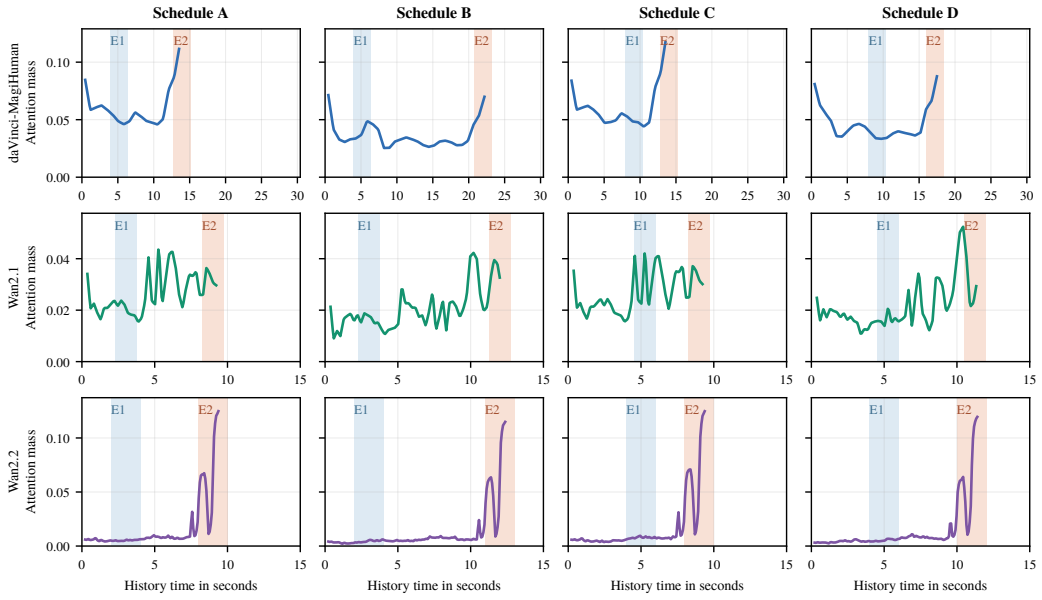

Figure 8: Historical KV readout for the cup event on the sparse-event test suite. Rows show the same three backbones as the phone-event result in the main text. The plots compare attention to the first event during the repeated event and during nearby idle periods. The four schedules place the same event at different temporal positions.

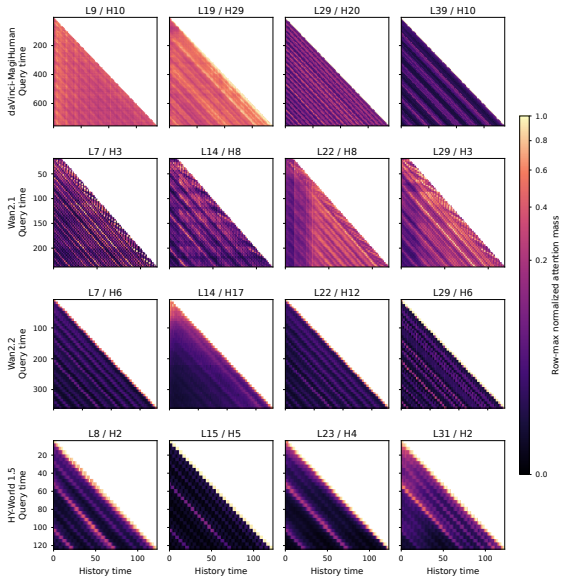

Figure 9: Additional layer- and head-specific historical KV readouts on the no-switch test suite. Rows correspond to backbones, and each column shows the layer and head named above the panel. Attention remains broadly distributed over middle history in the displayed results.

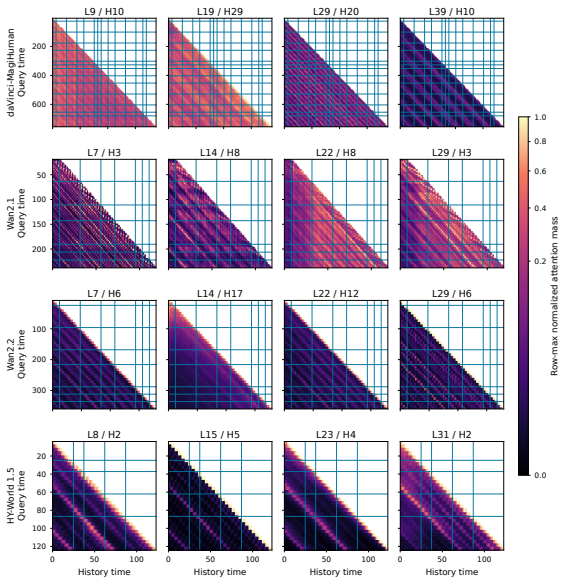

Figure 10: Additional layer- and head-specific historical KV readouts on the prompt-switch test suite. Rows correspond to backbones, and each column shows the layer and head named above the panel. Cyan lines mark prompt-switch boundaries. Attention remains broadly distributed over middle history rather than consistently concentrating at the switch boundaries.

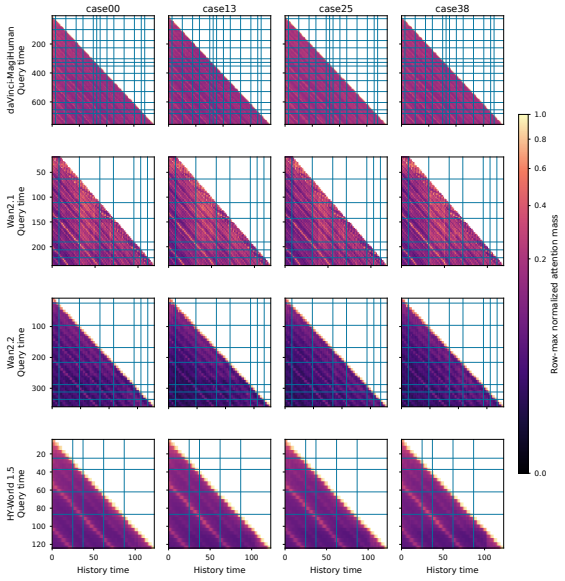

Figure 11: Additional case-level historical KV readouts on the prompt-switch test suite. Rows correspond to backbones, and columns show the same four cases (case00, case13, case25, and case38). Cyan lines mark prompt-switch boundaries. The displayed results retain the broad temporal patterns observed in the aggregate analysis.

Table 3: Long-horizon policy comparison on the no-switch test suite at 30s, 60s, and the longest evaluation endpoint for each backbone. All methods use the complete set of 39 cases. Uniform Reservoir and TWR are averaged over six fixed retention seeds within each case.

Backbone	Strategy	FaceSim↑	ID drift↓	Face det.↑	DINO↑	Scene drift↓	Image Q↑	Q drift↓	Dynamic	Motion↑
					30s					
daVinci-MagiHuman	sink+fifo	36.2	24.1	99.7	83.3	3.9	69.2	3.0	22.3	99.4
	Deep Forcing	38.9	20.7	99.7	84.2	3.3	69.1	2.2	19.9	99.4
	Uniform reservoir	42.8	8.8	99.8	85.5	1.1	69.2	1.6	17.1	**99.5**
	TWR	**44.7**	**6.0**	**99.9**	**86.0**	**0.6**	**69.4**	**1.2**	11.0	**99.5**
Wan2.1	sink+fifo	26.2	11.2	**100.0**	79.0	6.9	71.0	5.2	8.3	99.1
	Deep Forcing	24.6	15.4	99.2	76.4	11.0	67.8	12.7	4.8	**99.3**
	Uniform reservoir	27.7	7.1	**100.0**	81.2	2.6	72.1	2.4	7.8	99.2
	LongLive-RAG	27.5	8.0	**100.0**	80.9	3.2	71.2	3.4	7.9	99.2
	TWR	**28.9**	**4.5**	**100.0**	**82.1**	**1.2**	**72.5**	**1.7**	8.9	99.2
Wan2.2	sink+fifo	49.1	33.0	96.5	82.5	10.0	62.1	6.0	26.6	99.4
	Deep Forcing	45.8	47.0	92.4	78.4	11.8	55.8	20.5	51.8	98.9
	Uniform reservoir	54.3	19.5	98.8	85.6	3.8	**62.2**	3.1	19.4	99.4
	TWR	**58.1**	**10.6**	**99.7**	**87.4**	**1.4**	62.1	**2.3**	17.9	**99.5**
HY-World 1.5	sink+fifo	62.4	23.8	**98.9**	88.8	7.5	70.4	4.0	1.3	99.6
	Deep Forcing	63.8	23.5	98.5	89.3	7.8	70.6	3.3	2.7	99.6
	Uniform reservoir	66.0	17.0	98.5	90.5	4.8	**71.0**	2.5	2.2	99.6
	TWR	**67.5**	**14.4**	98.3	**91.1**	**3.9**	70.9	**2.4**	2.2	**99.7**
					60s					
daVinci-MagiHuman	sink+fifo	27.2	29.5	99.5	80.2	3.8	**69.7**	3.5	22.0	99.4
	Deep Forcing	29.2	28.9	99.5	80.9	3.6	69.5	3.2	21.0	99.4
	Uniform reservoir	40.9	8.1	99.9	85.2	1.0	69.5	2.1	11.8	99.5
	TWR	**44.1**	**4.4**	**100.0**	**86.0**	**0.5**	69.5	**1.4**	7.9	**99.6**
Wan2.1	sink+fifo	21.0	18.7	**100.0**	72.0	18.0	66.2	18.1	6.8	**99.2**
	Deep Forcing	18.5	22.5	93.5	64.3	23.0	59.2	25.6	3.3	98.9
	Uniform reservoir	25.8	7.6	**100.0**	79.5	5.0	71.3	3.6	5.8	**99.2**
	LongLive-RAG	24.4	11.9	**100.0**	77.5	9.7	69.9	6.3	6.4	**99.2**
	TWR	**28.1**	**3.7**	**100.0**	**81.5**	**1.7**	**72.4**	**1.9**	8.0	**99.2**
Wan2.2	sink+fifo	37.4	46.2	92.6	74.0	18.2	59.0	11.8	32.3	99.2
	Deep Forcing	34.8	61.7	68.2	60.3	27.4	43.4	35.7	56.0	98.4
	Uniform reservoir	47.2	26.7	98.7	83.0	5.6	60.8	5.0	17.4	99.3
	TWR	**55.7**	**10.8**	**99.8**	**87.0**	**1.6**	**61.6**	**2.8**	16.8	**99.4**
HY-World 1.5	sink+fifo	52.4	41.2	95.8	78.3	19.2	68.7	7.9	2.7	99.5
	Deep Forcing	53.9	45.6	92.8	78.1	21.8	68.8	7.3	5.5	99.3
	Uniform reservoir	59.8	29.2	95.3	86.0	10.0	**70.7**	3.4	3.8	99.5
	TWR	**62.6**	**21.4**	**97.1**	**88.2**	**7.8**	70.5	**3.3**	3.5	**99.6**
					240s					
daVinci-MagiHuman	sink+fifo	18.5	19.6	99.7	76.3	5.4	**70.0**	3.6	20.6	99.4
	Deep Forcing	18.8	23.1	99.6	76.4	5.3	69.9	3.9	21.7	99.4
	Uniform reservoir	34.9	12.5	**100.0**	83.8	2.0	69.8	2.4	5.9	**99.6**
	TWR	**42.1**	**4.4**	**100.0**	**85.7**	**0.8**	69.7	**1.5**	5.4	**99.6**
Wan2.1	sink+fifo	11.9	22.2	70.8	44.3	31.4	47.3	30.0	14.7	96.2
	Deep Forcing	13.7	22.8	38.7	31.8	36.0	47.6	29.2	26.1	93.7
	Uniform reservoir	22.2	7.5	**100.0**	75.0	9.8	69.3	5.8	4.5	**99.3**
	LongLive-RAG	14.6	23.5	82.3	54.7	26.8	57.1	28.9	11.3	97.0
	TWR	**26.8**	**2.9**	**100.0**	**80.5**	**2.5**	**71.7**	**2.6**	6.8	99.2
Wan2.2	sink+fifo	19.0	44.2	69.3	48.0	29.1	49.6	19.8	28.4	98.3
	Deep Forcing	28.2	38.4	25.2	26.2	32.6	32.2	33.7	32.6	96.7
	Uniform reservoir	30.1	33.3	97.9	76.2	8.9	57.5	8.9	15.0	99.1
	TWR	**42.9**	**24.1**	**99.9**	**84.6**	**3.6**	**60.6**	**3.7**	14.2	**99.4**
					120s					
HY-World 1.5	sink+fifo	44.3	64.6	72.6	57.7	38.0	65.7	11.5	10.0	99.0
	Deep Forcing	48.2	66.6	64.5	54.7	37.4	65.1	15.7	8.7	98.8
	Uniform reservoir	53.4	39.9	87.3	77.5	**17.0**	**69.8**	5.6	6.1	99.2
	TWR	**56.0**	**36.3**	**91.0**	**80.2**	**17.0**	69.7	**5.4**	5.4	**99.4**

E.2 Results on the Prompt-Switch Test Suite

Evaluation on the prompt-switch test suite measures how each cache policy handles changes in the requested action while preserving persistent visual content. Each method receives the same action sequence and switch times for a given case.

TWR preserves more identity information over long generation horizons in the reported comparison on the prompt-switch test suite. The result is reported together with content, motion, quality, and face coverage metrics.

Table 4: Policy comparison on the prompt-switch test suite at 30s, 60s, and the longest evaluation endpoint for each backbone. All methods use the complete set of 39 cases and the same frozen action schedule within each case. Uniform Reservoir and TWR are averaged over six fixed retention seeds.

Backbone	Strategy	FaceSim↑	ID drift↓	Face det.↑	DINO↑	Scene drift↓	Image Q↑	Q drift↓	Dynamic	Motion↑
					30s					
daVinci-MagiHuman	sink+fifo	35.0	25.5	99.6	81.9	4.8	**69.4**	3.3	30.8	99.4
	Deep Forcing	36.5	24.9	99.6	82.8	4.1	69.2	2.3	29.2	99.4
	Uniform reservoir	41.4	11.7	99.6	84.9	1.6	69.2	1.8	21.3	**99.5**
	TWR	**43.8**	**7.6**	**99.9**	**85.6**	**0.8**	**69.4**	**1.3**	13.5	**99.5**
Wan2.1	sink+fifo	26.2	10.8	**100.0**	79.1	6.7	71.2	4.7	7.5	99.1
	Deep Forcing	25.5	13.1	99.9	76.5	10.7	68.5	10.9	4.8	**99.3**
	Uniform reservoir	27.9	6.6	**100.0**	81.2	2.6	72.3	2.3	7.9	99.2
	LongLive-RAG	27.6	7.8	**100.0**	80.7	3.4	71.5	2.8	6.4	99.2
	TWR	**28.8**	**4.1**	**100.0**	**81.9**	**1.2**	**72.7**	**1.7**	9.0	**99.3**
Wan2.2	sink+fifo	47.5	36.2	97.6	82.8	9.7	61.0	6.1	26.1	99.4
	Deep Forcing	43.3	47.2	93.7	78.5	11.4	56.0	19.7	47.9	98.9
	Uniform reservoir	53.2	21.1	98.5	85.7	3.7	61.9	3.7	18.8	99.4
	TWR	**57.5**	**11.2**	**99.6**	**87.5**	**0.9**	**62.0**	**2.3**	17.6	**99.5**
HY-World 1.5	sink+fifo	62.8	23.0	98.5	88.7	7.7	70.5	3.6	1.8	**99.6**
	Deep Forcing	64.3	22.3	**98.7**	89.1	8.1	70.7	2.9	1.3	**99.6**
	Uniform reservoir	65.9	17.6	98.6	90.5	4.8	**71.1**	**2.4**	2.2	**99.6**
	TWR	**67.3**	**14.8**	98.3	**91.1**	**4.0**	70.9	2.8	2.5	**99.6**
					60s					
daVinci-MagiHuman	sink+fifo	25.4	31.9	99.4	78.4	5.0	**69.8**	3.7	32.5	99.4
	Deep Forcing	26.1	32.9	99.2	78.9	4.3	69.6	3.5	33.0	99.4
	Uniform reservoir	38.6	12.1	99.7	84.3	1.6	69.5	2.2	18.8	**99.5**
	TWR	**42.8**	**5.4**	**99.9**	**85.5**	**0.7**	69.6	**1.5**	11.1	**99.5**
Wan2.1	sink+fifo	20.9	18.5	**100.0**	72.2	17.7	67.2	16.0	6.9	99.1
	Deep Forcing	19.4	21.7	95.3	64.4	22.9	60.2	23.5	2.6	98.9
	Uniform reservoir	26.0	7.3	**100.0**	79.6	4.8	71.6	3.3	5.6	99.2
	LongLive-RAG	24.8	10.4	**100.0**	77.7	9.0	70.4	5.3	4.2	99.2
	TWR	**28.1**	**3.5**	**100.0**	**81.4**	**1.7**	**72.6**	**1.9**	8.1	**99.3**
Wan2.2	sink+fifo	34.9	49.0	92.9	75.1	17.9	57.7	10.9	31.8	99.2
	Deep Forcing	33.2	61.4	67.3	59.5	26.3	43.8	35.6	55.3	98.4
	Uniform reservoir	45.5	27.7	98.4	83.3	5.7	60.4	5.5	17.4	99.3
	TWR	**54.9**	**11.2**	**99.7**	**87.1**	**1.6**	**61.5**	**3.1**	16.4	**99.5**
HY-World 1.5	sink+fifo	52.3	39.1	**97.1**	79.7	18.5	68.8	6.5	1.9	99.5
	Deep Forcing	52.7	48.7	93.1	77.1	23.5	68.8	6.7	5.4	99.3
	Uniform reservoir	59.5	30.7	94.8	85.7	10.3	**70.7**	**3.5**	4.4	99.5
	TWR	**62.2**	**22.4**	96.5	**88.0**	**8.3**	70.6	**3.5**	4.7	**99.6**
					240s					
daVinci-MagiHuman	sink+fifo	16.1	20.8	99.4	74.2	6.3	**70.1**	4.1	32.0	99.4
	Deep Forcing	16.0	21.8	99.4	74.2	6.1	**70.1**	4.1	34.0	99.4
	Uniform reservoir	30.7	16.2	99.7	82.1	2.7	69.8	2.6	13.4	99.5
	TWR	**40.4**	**5.5**	**99.9**	**85.1**	**1.1**	69.8	**1.6**	8.9	**99.6**
Wan2.1	sink+fifo	10.8	21.9	81.5	47.8	29.5	48.6	32.2	14.8	95.8
	Deep Forcing	14.7	23.6	39.4	33.4	36.5	51.1	23.8	18.2	95.2
	Uniform reservoir	22.6	7.2	**100.0**	75.5	9.4	69.8	5.4	2.8	**99.3**
	LongLive-RAG	15.3	21.2	86.5	57.4	23.2	59.2	22.8	14.9	97.1
	TWR	**26.9**	**2.7**	**100.0**	**80.5**	**2.4**	**72.1**	**2.3**	6.7	**99.3**
Wan2.2	sink+fifo	15.6	41.5	74.7	50.8	27.9	48.7	21.3	19.1	98.7
	Deep Forcing	25.6	36.5	30.4	27.4	31.9	30.4	36.1	50.5	93.8
	Uniform reservoir	27.9	33.9	97.4	76.6	8.8	56.7	9.4	13.6	99.0
	TWR	**41.8**	**25.0**	**99.6**	**84.8**	**3.5**	**60.5**	**3.9**	13.8	**99.3**
					120s					
HY-World 1.5	sink+fifo	43.6	65.6	72.6	58.8	37.3	66.2	12.3	8.3	99.0
	Deep Forcing	46.3	68.6	63.2	54.0	38.1	65.2	11.9	9.5	98.7
	Uniform reservoir	53.0	41.4	86.3	76.6	18.2	69.7	6.0	7.0	99.2
	TWR	**56.5**	**34.5**	**89.6**	**79.8**	**17.2**	**69.9**	**4.9**	6.7	**99.4**

F CONTROLLED HISTORY INTERVENTIONS

We use controlled interventions to test whether retained historical KV entries affect later generation. Each intervention changes a specified part of the retained history while keeping the initial frame, prompt sequence, generation seed, and duration fixed.

Each intervention run is compared with a matched control run. The difference between the two runs measures the effect of the modified historical content.

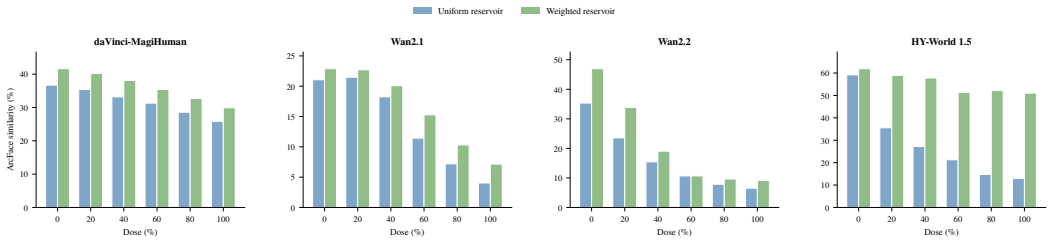

Figure 12: Effect of the face swap history intervention on later identity similarity. Each bar reports full-video ArcFace similarity at one corruption dose under matched generation settings. Planned samples without a valid target-face measurement contribute zero, so this diagnostic also captures losses in face-detection coverage.

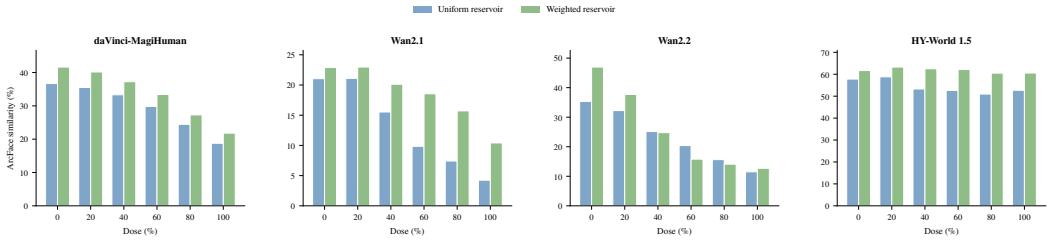

Figure 13: Effect of the horizontal strip history intervention on later identity similarity. Each bar reports full-video ArcFace similarity at one corruption dose under matched generation settings. Planned samples without a valid target-face measurement contribute zero, so this diagnostic also captures losses in face-detection coverage.

F.1 Face Swap Intervention

The face swap intervention inserts identity-related content from another case into the retained history. We then measure identity similarity in later generated frames.

The intervention changes later identity measurements relative to the matched control. This result shows that identity information stored in historical KV entries can affect later generation.

F.2 Horizontal Strip Intervention

The horizontal strip intervention permutes five fixed full-width bands in selected middle-history latent states before their KV entries are recomputed. The dose controls how many middle states are modified, with larger doses using nested sets. The sink, recent window, current block, and unselected states remain unchanged.

The displaced spatial pattern affects later frames, and ArcFace similarity drops at higher corruption doses. This shows that spatial information stored in middle KV can influence later visual content.