KEEPSAKE: SELECTIVE SPATIAL MEMORY FOR LONG-HORIZON VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon camera-controlled video generation relies on persistent memory to maintain scene consistency beyond the model's immediate temporal context. Existing systems follow two strategies to achieve this consistency. Full-history approaches retain all generated observations, causing unbounded storage and retrieval costs to grow while allowing low-fidelity memories to persist. Selective-construction approaches reduce redundancy, but make one-time retention decisions that are never revisited, even as an observation's value changes with the evolving memory bank. Both strategies leave a shared question unresolved: as the generated history evolves, which stored observations should still remain in memory? Our key insight is that the value of a stored observation is not fixed, but relational: it depends on the alternatives currently available in the memory bank. A view supported by many geometrically and visually similar substitutes can be relinquished with little loss of coverage, whereas an observation with few viable alternatives should remain regardless of age. We introduce KEEPSAKE, an online, training-free controller for fixed-capacity spatial memory. At each update, KEEPSAKE constructs a pose–appearance graph over retained and newly generated observations, combining camera-pose proximity with visual similarity. A retention priority jointly captures the number of strong substitutes and the similarity of the closest alternative, allowing KEEPSAKE to continually reassess memory value, preserve observations with little alternative support, and evict highly replaceable ones under a fixed budget. The controller modifies only the persistent-memory update; the host generator, denoising schedule, and retrieval rule remain unchanged. Across MemCam and WorldMem, KEEPSAKE improves FVD and LPIPS under a fixed memory budget; on 180-second MemCam trajectories, it retains only 32 of 5,397 frames while reducing FVD by 35.1%.

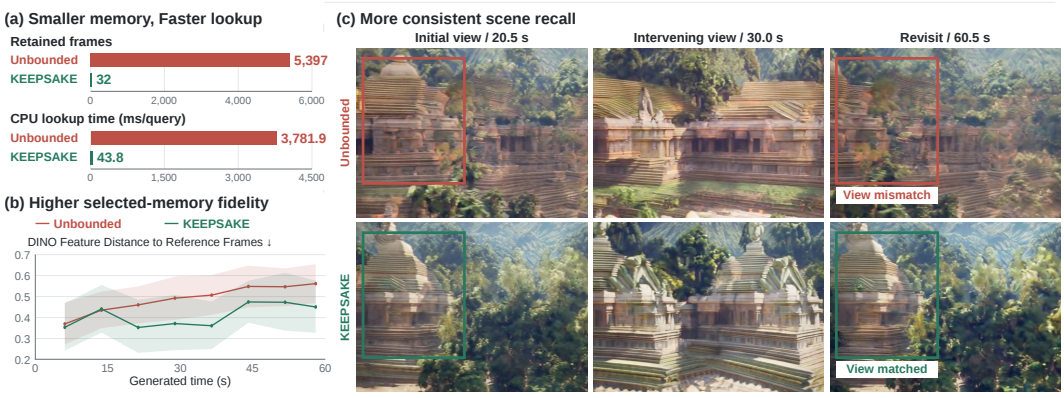


Figure 1: **What should remain in memory?** Unbounded memory versus our budgeted memory in MemCam. (a) Retained-frame count and CPU lookup latency per query. (b) DINO feature cosine distance from selected historical frames to reference frames over generation time; lower is better. (c) A camera-revisit example, with boxes highlighting corresponding scene regions.

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

1 INTRODUCTION

Camera-controlled video generation enables deliberate viewpoint navigation within generated scenes, supporting interactive exploration and controllable content creation. Extending this capability to long horizons requires preserving scene consistency beyond the model's immediate temporal context, particularly when previously observed regions are revisited after substantial temporal gaps. Recent memory-augmented methods address this challenge by maintaining historical observations associated with camera states, retrieving relevant evidence for subsequent generation, and writing newly generated observations back into persistent memory Gao et al. (2026); Xiao et al. (2026); Yu et al. (2025); Li et al. (2025). Each generation step is therefore both an output and a potential source of evidence for future generation.

Existing spatial-memory systems commonly retain the full generation history and rely on retrieval to select relevant observations at query time Gao et al. (2026); Xiao et al. (2026). This strategy introduces two coupled limitations. First, the memory archive grows continuously with generation, leading to increasing retrieval overhead as the number of stored observations expands (Fig. 1(a)). Second, quality degradation can persist in the memory archive, allowing drifted observations to be repeatedly retrieved and reused. As shown in Fig. 1(b), memories with well-matched viewpoints can still deviate substantially from the established scene, indicating that geometric relevance alone does not guarantee reliable memory. Fig. 1(c) further shows the resulting inconsistency when the camera revisits the same region, where the regenerated scene no longer matches its earlier appearance.

Selective-construction approaches reduce redundant accumulation through spatial admission, keyframe selection, or memory merging Li et al. (2025); Wang et al. (2026); Yu et al. (2026). However, these decisions are typically made only when an observation is written and are not revisited as the memory evolves. This is limiting because the value of a stored observation is inherently dynamic: a view that is unique at admission may later become redundant as similar evidence accumulates, while an older view with few substitutes may remain important. This raises a shared question across existing memory strategies: as generation proceeds, which observations should still remain in memory?

We formulate this as a dynamic reassessment problem. The retention value of an observation is relational rather than intrinsic, depending on the alternatives currently available in memory. A view with several geometrically and visually similar substitutes can be removed with limited loss of coverage, whereas an observation with few viable alternatives should be retained regardless of its age or admission-time novelty. Selective forgetting therefore serves not only to bound memory growth, but also to continually reallocate capacity toward observations that remain most informative.

We introduce KEEPSAKE, an online, training-free, plug-and-play memory controller that dynamically reassesses which observations should remain in memory. At each update, KEEPSAKE pools the retained memory with newly generated observations and constructs a pose–appearance graph whose edge weights combine camera-pose proximity with visual similarity. Each observation is scored by its number of strong substitutes and the similarity of its closest alternative. Low-priority observations, those with many strong substitutes are evicted until a fixed budget B is satisfied. Crucially, retention priorities depend on the current candidate set rather than being fixed at admission, so an observation that was previously important can become replaceable as new alternatives arrive. The fixed budget bounds memory growth and retrieval overhead, while dynamic reassessment continually reallocates capacity toward observations that remain most useful.

Our experiments show selective forgetting can improve generation quality while substantially reducing memory size. On fifteen matched 180-second MemCam Gao et al. (2026) trajectories, KEEPSAKE retains only 32 observations instead of 5,397 while reducing FVD from 734.2 to 476.6, a 35.1% improvement. On fifteen matched 60-second WorldMem Xiao et al. (2026) videos, FVD decreases from 3077.6 to 1116.9. Importantly, simply imposing a memory cap is not sufficient: FIFO retention degrades generation quality relative to full retention, and different bounded policies yield substantially different results under the same budget (Tab. 1). These results show memory quality depends not only on how much history is retained, but also on which observations remain available.

Our contributions can be summarized as follows:

- We identify a fundamental limitation of growing spatial memory in camera-controlled long-horizon video generation and formulate retention as a dynamic reassessment problem: an

observation's value depends on the substitutes currently available in memory, rather than its age or admission-time novelty.

- We propose KEEPSAKE, a training-free memory controller that continuously reassesses retention value through pose–appearance substitutability, preserving observations with few viable alternatives under a fixed budget.
- Experiments on MemCam and WorldMem show that selective forgetting can simultaneously improve long-horizon generation quality and substantially reduce memory and retrieval cost, outperforming both full-history retention and bounded baselines.

2 RELATED WORK

Camera-Controlled Video Generation. Recent text-to-video and image-to-video models have achieved impressive visual quality and realism Chen et al. (2024); Yang et al. (2025); Kong et al. (2024); Blattmann et al. (2023). However, they offer limited control over viewpoint evolution. Explicit camera control is therefore essential for applications that require deliberate viewpoint navigation and interactive exploration of a generated scene. Accordingly, recent studies directly condition video generation on camera trajectories Wang et al. (2024); Yang et al. (2024); He et al. (2024); Zheng et al. (2024); Hou & Chen (2024). Yang et al. (2024); Wang et al. (2024) disentangle camera motion from object motion, while He et al. (2024) introduce explicit camera-pose conditioning into pretrained video generators. More recent work further improves camera controllability through geometric and correspondence priors, including epipolar constraints, trajectory-based correspondence, monocular depth, and explicit 3D scene representations Zheng et al. (2024); Hou & Chen (2024); Xiao et al. (2025); Li et al.; Ren et al. (2025). Building on camera-controllable video generation, KEEPSAKE focuses on maintaining scene consistency over long horizons, particularly preserving previously observed content when the camera revisits earlier viewpoints after extended exploration.

Long-Horizon Consistency under Camera Control. Long-horizon camera exploration requires preserving scene information beyond the model's local context, particularly when previously observed regions are revisited after large temporal or viewpoint gaps. Recent methods address this challenge with persistent spatial memory, storing historical observations and retrieving evidence relevant to the current viewpoint Xiao et al. (2026); Yu et al. (2025); Gao et al. (2026). Other approaches organize memory through geometry-aware indexing, spatial correspondence, local spatial memories, or persistent 3D representations to better preserve cross-view consistency Li et al. (2025); Ren et al. (2025); Xu et al. (2026); Zhang et al. (2026); Zhao et al.; Wang et al. (2026); Yu et al. (2026); Hong et al. (2025). Beyond full-history memory, long-video generation has also explored adaptive memory selection and bounded retention under limited context budgets Ji et al. (2025); Shi et al. (2026); Li et al. (2026). However, these methods either retain an expanding archive or make retention decisions primarily at admission time. KEEPSAKE further studies how memory should be maintained as it grows, showing that retaining more historical observations is not always beneficial. It selectively preserves observations that remain useful for future revisits, improving revisit consistency while controlling storage and computation.

3 KEEPSAKE: MEMORY BUDGETING METHOD

KEEPSAKE maintains a memory bank of at most B observations by reassessing redundancy as new views arrive. The retention priority of an image depends on the alternatives currently available: how many views provide similar evidence, and how closely its strongest alternative matches it. We capture these relationships through a pose-appearance graph that combines camera geometry with visual similarity. At each update, the controller ranks retained and incoming images, favoring views with limited alternative support and removing low-priority observations to satisfy the budget. We pose memory retention as a prospective policy, as in interactive video world models, there is no certainty on where the camera will move next. Figure 2 illustrates this update within the existing generation loop; the generator and retrieval rule remain unchanged. In this section, we cover how our strict memory budgeting is enforced and updated at each step along with construction of the pose-appearance graph.

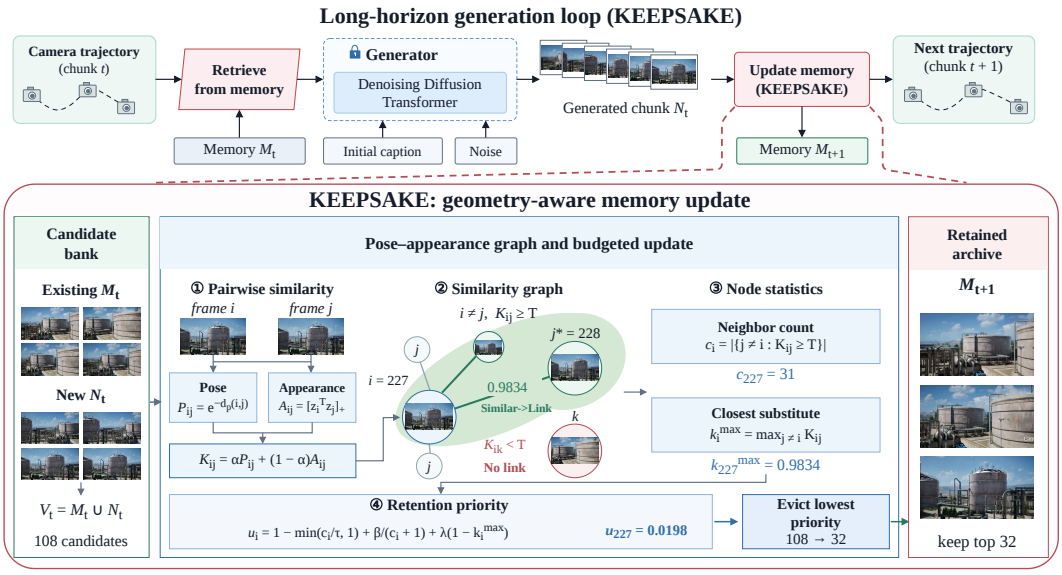


Figure 2: **Overview of KEEPSAKE. Top:** KEEPSAKE is inserted into the long-horizon generation loop as a memory-update module. After each generated chunk, it updates the persistent memory while leaving the generator and retrieval rule unchanged. **Bottom:** Existing memory and newly generated observations are pooled into a candidate bank and organized as a pose-appearance graph, where edge weights combine camera-pose proximity and visual similarity. Node statistics capture both the number of strong substitutes and the similarity to the closest substitute, which jointly determine retention priority. Low-priority observations are evicted until the fixed budget B is met. Thumbnails and displayed scores are illustrative.

3.1 FIXED-BUDGET SPATIAL MEMORY

At step t, the generator produces a video chunk $\hat{Y}_t$ along a prescribed camera trajectory P_t. Let M_{t-1} denote the memory bank available before generation, and let B be its capacity in stored observations. The retriever R selects historical context C_t according to the upcoming camera poses. The generator G then synthesizes the next chunk using this context and the preceding chunk $\hat{Y}_{t-1}$:

$$C_t = R(P_t; M_{t-1}), \qquad \hat{Y}_t = G(\hat{Y}_{t-1}, P_t, C_t). \tag{1}$$

After generation, the host constructs N_t, the set of new memory observations obtained from $\hat{Y}_t$. In the image-memory setting, each observation contains a generated frame, its camera pose, and an appearance descriptor. These observations are combined with the existing memory bank to form the candidate set $V_t = M_{t-1} \cup N_t$. KEEPSAKE determines which candidates remain:

$$M_t = U(M_{t-1}, N_t; B), \qquad M_t \subseteq V_t, \qquad |M_t| \leq B. \tag{2}$$

Only the update rule U changes; the generator and retriever remain unchanged. The budget limits the persistent memory bank, rather than merely the context retrieved for a single generation step.

3.2 POSE-APPEARANCE GRAPH

We construct a graph whose nodes are the candidate observations in V_t. Each observation o_i is associated with a camera-to-world pose $(R_i, \mathbf{t}_i)$ and a normalized visual descriptor $\mathbf{z}_i$. Here, R_i and $\mathbf{t}_i$ denote camera rotation and position, respectively. The edge weight $K(i,j)$ combines camera-pose proximity with descriptor similarity: the former measures closeness in viewing position and direction, while the latter captures visual resemblance. Together, these signals provide a proxy for shared view coverage, which we use to assess redundancy within the memory bank.

Pose distance. We measure camera-pose separation by combining translation and rotation:

$$d_p(i,j) = \frac{\|\mathbf{t}_i - \mathbf{t}_j\|_2}{\max(\widetilde{d}_t, \epsilon)} + \frac{2\theta_{ij}}{\pi}, \tag{3}$$

where $\widetilde{d}_t$ is the median nonzero pairwise translation distance among observations in V_t, and $\theta_{ij} \in [0, \pi]$ is the rotation angle of $R_i^\top R_j$. The translation normalization expresses positional separation relative to the current candidate set, while the rotation term accounts for differences in viewing direction. We use $\widetilde{d}_t = 1$ when all camera positions coincide and $\epsilon = 10^{-6}$ as a numerical floor.

Edge weights. We combine pose proximity and appearance similarity to define the weight between distinct observations:

$$K(i,j) = \alpha\, e^{-d_p(i,j)} + (1-\alpha)\max(\mathbf{z}_i^\top \mathbf{z}_j, 0), \qquad i \neq j, \tag{4}$$

where $\alpha \in [0,1]$ controls the relative contribution of geometry and appearance. The exponential converts pose distance into similarity. Since the visual descriptors are normalized, their inner product measures cosine similarity, with negative values clipped to zero. The resulting weights are symmetric and lie in $[0,1]$. Larger weights indicate greater similarity between observations, providing the basis for measuring redundancy within the memory bank.

Node degree and maximum similarity. To assess redundancy, we compute two statistics for each observation: its degree in the thresholded graph, which counts strongly similar observations, and its maximum similarity to any other candidate. Thresholding the edge weights at $T \in [0,1]$ defines an undirected graph $\mathcal{G}_t = (V_t, E_t)$, with

$$E_t = \{\{o_i, o_j\} : i < j,\ K(i,j) \geq T\}\,. \tag{5}$$

For each observation, we compute its degree in this graph and the weight of its strongest alternative:

$$c_i = \sum_{j \neq i} \mathbb{I}[K(i,j) \geq T], \qquad k_i^{\max} = \max_{j \neq i} K(i,j). \tag{6}$$

The degree c_i counts observations with high geometric–visual similarity, while $k_i^{\max}$ measures the similarity of the closest alternative. These statistics distinguish how many strong alternatives are available from how closely the best one matches the observation. The maximum is computed over all other candidates in V_t, including those whose weights fall below T.

3.3 Retention Priority

We combine node degree and maximum similarity into a retention priority. Low-degree observations have fewer strongly similar alternatives, while a large maximum similarity indicates a close match elsewhere in the candidate set. We therefore assign:

$$\begin{aligned} u_i &= \phi(c_i) + \lambda(1 - k_i^{\max}), \\ \phi(c) &= 1 - \min\Big(\frac{c}{\tau}, 1\Big) + \frac{\beta}{c+1}, \end{aligned} \tag{7}$$

where higher u_i indicates greater retention priority. The degree-based function ϕ decreases as an observation acquires more strongly similar alternatives. Its linear term reaches zero at degree τ, while the inverse-degree term continues to distinguish nodes beyond this threshold. The remaining term, $1 - k_i^{\max}$, increases priority when even the closest alternative is dissimilar. The parameters $\beta \geq 0$ and $\lambda \geq 0$ control the contributions of inverse degree and maximum similarity, respectively, with $\tau > 0$ specifying the degree threshold.

Retention priorities depend on the current candidate set V_t, rather than being fixed when an observation first enters memory. New observations can provide additional alternatives for an existing view, reducing its priority, while older views with few alternatives remain favored. When the candidate set exceeds the budget, KEEPSAKE computes all priorities once and uses this fixed ranking throughout eviction. Node degrees and maximum similarities are not recomputed after individual removals.

3.4 Budgeted Memory Update

At each update, KEEPSAKE evaluates retention priorities over the candidate set $V_t = M_{t-1} \cup N_t$. When $|V_t| > B$, it removes the $|V_t| - B$ lowest-priority observations eligible for eviction, breaking equal-priority ties by earlier insertion. Otherwise, all observations are retained. Priorities are computed once per update and remain fixed during removal in Algorithm 1. The resulting memory bank supplies candidates to the original retriever, without changing its ranking rule.

Algorithm 1 KEEPSAKE: budgeted memory update

Require: Memory bank M_{t-1}, new observations N_t, budget B
Ensure: Updated memory bank M_t, with $|M_t| \leq B$
$V_t \leftarrow M_{t-1} \cup N_t$
if $|V_t| \leq B$ **then**
  return V_t
end if
Compute edge weights $K(i, j)$ over V_t using Eq. (4)
Compute degrees c_i and maximum similarities $k_i^{\max}$ using Eq. (6)
Compute retention priorities u_i once using Eq. (7)
Rank eviction-eligible observations by increasing u_i, breaking ties by earlier insertion
$M_t \leftarrow V_t$
Remove the first $|V_t| - B$ ranked observations from M_t
return M_t

3.5 COMPUTATIONAL COST

Let s denote the maximum number of new observations introduced per chunk. Since KEEPSAKE retains at most B observations, each update considers at most $B+s$ candidates. Computing pairwise pose–appearance similarities and retention priorities requires $O((B+s)^2)$ time, while sorting adds $O((B+s)\log(B+s))$. The retained archive uses $O(B)$ storage, and exhaustive retrieval scores at most B candidates per query. Therefore, for fixed B and chunk size, memory usage, update cost, and retrieval cost remain bounded independently of video duration. KEEPSAKE modifies only the persistent-memory update and requires no generator retraining or changes to the retrieval rule.

4 EXPERIMENTS

We investigate how much spatial memory long-horizon generation needs and how that capacity should be allocated. Our primary comparisons use MemCam and WorldMem, varying memory retention while keeping the generator and retrieval rule unchanged within each system. Comparisons against complete retention test whether a compact memory bank can preserve generation quality; comparisons at matched budgets test whether the choice of retained observations matters.

4.1 EXPERIMENTAL SETUP

Generation systems and datasets. We evaluate KEEPSAKE in two settings: MemCam (Gao et al., 2026) on the Context-as-Memory dataset (Yu et al., 2025), and WorldMem (Xiao et al., 2026) on its Minecraft world-simulation dataset. Each primary evaluation contains fifteen matched 60-second videos. A separate set of fifteen matched 180-second MemCam videos also test longer generation. Within each system, retention policies share the initial inputs, prescribed control sequences, generator checkpoint, generation settings, and random seeds. We report comparisons separately for each system and duration, so that differences between retention policies are measured under the same generation conditions.

Retention baselines. We compare KEEPSAKE with complete retention and three bounded baselines: FIFO, K-center, and Marginal Coverage Eviction (MCE). FIFO prioritizes recency by discarding the oldest eligible observations. K-center selects representatives to improve worst-case coverage, while MCE removes observations whose deletion minimally reduces coverage over representative historical queries. Both coverage-based baselines use geometry and appearance, whereas KEEPSAKE ranks observations by redundancy within the current candidate bank using node degree and maximum similarity. All bounded methods use the same archive budget, with $B = 32$ in the main comparison. Additional algorithmic details are provided in the appendix B.1.

Implementation details. We use MemCam with the Wan2.1 1.3B Wan et al. (2025) backbone, generating 76-frame sections with 50 denoising steps. For retention scoring, generated images are represented by ℓ_2-normalized DINOv2-Base Oquab et al. (2023) descriptors.

Table 1: Main quality results on 15 matched 60-second videos within each system. All bounded methods use $B = 32$. For consistency, we report FVD and LPIPS. For efficiency, we report number of stored items in memory and retrieval time needed for the model for each generation chunk. For additional evaluation we also report Vbench scores across subject and background consistency, as well as aesthetics and average scores.

	Consistency		VBench ↑				Efficiency	
Models	FVD ↓	LPIPS ↓	Subject	Background	Aesthetic	Average	Stored items ↓	Retrieval ms/query ↓
MemCam	784.9	0.59	80.53	90.16	45.42	73.98	1825	1344.61
MemCam + FIFO	832.2	0.62	77.01	89.50	45.35	73.34	32	**39.93**
MemCam + MCE	773.9	0.60	79.78	89.97	46.05	73.96	32	54.20
MemCam + K-center	712.0	0.59	80.85	90.00	45.52	74.11	32	43.59
MemCam + **KEEPSAKE (Ours)**	**690.5**	**0.58**	**81.78**	**90.47**	**46.14**	**74.65**	32	43.77
WorldMem	3077.6	0.65	73.87	87.53	38.37	68.66	1,200	77.92
WorldMem + FIFO	3554.9	0.69	72.47	86.15	36.92	68.61	32	4.670
WorldMem + MCE	1924.0	0.57	79.12	90.30	41.54	72.67	32	4.71
WorldMem + K-center	1419.6	0.56	79.60	91.58	43.04	72.84	32	5.00
WorldMem + **KEEPSAKE (Ours)**	**1116.9**	**0.53**	**81.71**	**92.95**	**43.72**	**74.05**	32	**4.54**

We use WorldMem's released Minecraft diffusion and VAE checkpoints, generating frames at 128×128 resolution with 20 DDIM steps and $\eta = 0$. For retention scoring, predicted VAE latents are spatially average-pooled to 4×4, flattened, and ℓ_2-normalized.

Both systems share the default retention parameters: $\alpha = 0.65$ for pose-appearance weighting, $T = 0.65$ for edge construction, and $(\tau, \beta, \lambda) = (3, 0.5, 0.25)$ for retention priority. We evaluate budgets $B \in \{16, 32, 64, 128\}$, using $B = 32$ for the main comparison.

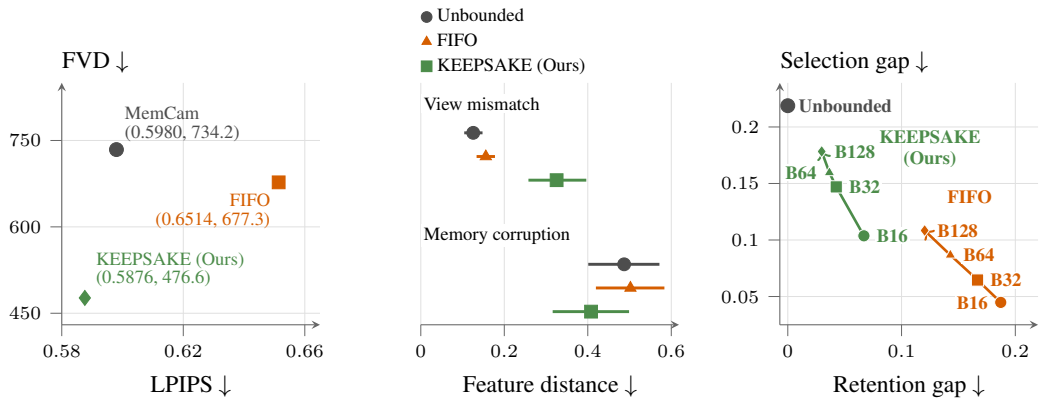
Evaluation metrics. We measure frame-level fidelity using LPIPS (Zhang et al., 2018), comparing generated frames with ground truth at the corresponding trajectory indices, and video-level distributional quality using FVD (Unterthiner et al., 2018). Lower values indicate better performance for both metrics. We additionally evaluate six VBench (Huang et al., 2024) dimensions: subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. These dimensions characterize perceptual quality and temporal behavior; we report them individually on a $\times 100$ scale rather than constructing an aggregate score.

4.2 Generation Quality under Bounded Memory

Table 1 compares complete retention with bounded-memory policies on the 60-second evaluation suites. With $B = 32$, KEEPSAKE improves both LPIPS and FVD over complete retention in both systems. In MemCam, LPIPS decreases from 0.5916 to 0.5850 and FVD from 784.9 to 690.5. In WorldMem, the corresponding improvements are 0.6523 to 0.5337 and 3077.6 to 1116.9. Among budget-matched policies, KEEPSAKE achieves the lowest LPIPS in both systems and the lowest FVD in WorldMem. FIFO, in contrast, worsens both metrics relative to complete retention in both systems. These results show that compact memory can improve generation quality, but imposing a capacity limit alone is insufficient: different retention rules yield substantially different outcomes under the same budget.

MemCam benchmark. We also evaluate KEEPSAKE using MemCam's original camera round-trip protocol on Context-as-Memory (Yu et al., 2025), covering 90° (153-frame) and 360° (609-frame) trajectories. We include Diffusion Forcing(DFoT) model Song et al. (2025) and Geometry Forcing Wu et al. (2026) in the comparison as well. As shown in Table 2, KEEPSAKE remains competitive with a memory budget of $B = 32$, achieving higher SSIM and lower LPIPS than MemCam in both settings. On the 360° round trip, PSNR also increases from 14.81 to 16.22 dB.

Longer rollouts. We evaluate fifteen matched 180-second MemCam trajectories to test bounded retention over a longer generation horizon, keeping the memory budget fixed at $B = 32$. Figure 3a reports this evaluation separately from the 60-second suite. Complete retention stores 5,397 ob-



(a) Long-horizon generation quality.

(b) View relevance and memory fidelity.

(c) Retention-selection tradeoff.

Figure 3: **Generation quality and memory diagnostics.** (a) LPIPS and FVD on fifteen matched 180-second MemCam rollouts. (b) View mismatch and memory corruption on fifteen matched 60-second rollouts, measured by DINOv2 cosine distance; whiskers denote 95% trajectory-bootstrap confidence intervals. FIFO and KEEPSAKE use $B = 32$ in (a,b). (c) Retention gap measures useful evidence lost through eviction, while selection gap measures the retriever's failure to select the best evidence that remains in memory. Results are shown on fifteen 180-second trajectories across memory budgets B. Lower is better for all metrics.

Table 2: **Round-trip and beyond-context evaluation.** KEEPSAKE improves most consistency metrics, with larger gains on the more challenging 360° round-trip setting.

	90°, 153 frames			360°, 609 frames		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DFoT	16.76	0.47	0.39	8.94	0.25	0.61
GeometryForcing	16.57	0.49	0.35	10.07	0.40	0.57
MemCam	**17.83**	0.51	0.357	14.81	0.42	0.50
KEEPSAKE	17.66	**0.52**	**0.34**	**16.22**	**0.46**	**0.43**

servations by the end of generation, whereas KEEPSAKE retains only 32. Despite this reduction, KEEPSAKE lowers FVD from 734.2 to 476.6 (35.1%). The improvement over complete retention therefore extends to three-minute generation without increasing the memory budget. Visual examples in figure 4 shows camera revisits maintain consistency with our model.

We restrict the 180-second extension to MemCam. In our setup, the original WorldMem configuration does not reliably sustain coherent generation beyond 60 seconds, exhibiting severe visual degradation. We therefore retain the 60-second prediction horizon as the stress-test setting for WorldMem and WorldMem + KEEPSAKE.

4.3 BUDGET SENSITIVITY AND RESOURCE SCALING

Quality across budgets. Table 3 reports the available capacity sweeps without mixing evaluation horizons. In the 180-second MemCam suite, the lowest reported FVD occurs at $B = 16$, while the lowest LPIPS occurs at $B = 64$. WorldMem's reported 60-second LPIPS is lowest at $B = 16$. Increasing capacity therefore does not consistently improve these within-policy scores. The sweeps characterize a quality–capacity tradeoff; they do not establish a universal best budget or an intrinsic penalty for candidate count.

Archive growth and retrieval workload. Under complete retention, the candidate pool grows continuously with generation. In MemCam, the mean unbounded candidate count increases from

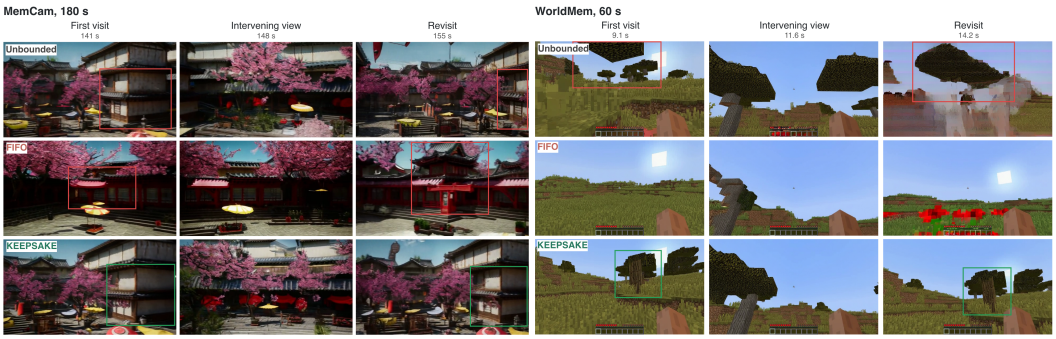


Figure 4: **Qualitative comparison of camera revisits.** Examples from 180-second MemCam rollouts (left) and 60-second WorldMem rollouts (right). Rows compare unbounded retention, FIFO, and KEEPSAKE with $B = 32$; columns show the first visit, an intervening view, and the revisit. Highlighted regions track scene content across visits. KEEPSAKE better preserves the appearance and structure of previously observed regions when the camera returns.

671.9 in the first 45 seconds to 4,689.5 in the final 45 seconds of a 180-second rollout. In contrast, KEEPSAKE caps the retained archive at 32 observations after each update while preserving the original retrieval interface and the amount of context supplied to the generator. This substantially reduces retrieval workload as generation proceeds. End-to-end generation time is reported in Table 6.

4.4 RETENTION AND RETRIEVAL DIAGNOSTICS

Retrieved-memory quality. Figure 3b shows that better viewpoint alignment does not necessarily imply more reliable historical evidence. Selected memories can match the query geometry while exhibiting greater visual drift, indicating that view relevance and memory fidelity are distinct. This distinction is important in long-horizon generation, because a geometrically suitable memory may still propagate degraded scene content. These results motivate evaluating not only which observations are stored, but also the fidelity of the evidence eventually retrieved from them.

Table 3: **Memory-budget ablation study.**

Budget B	LPIPS ↓	FVD ↓	VBench ↑	Query time ↓
MemCam, 180 seconds				
16	0.5876	**446.1**	**74.90**	**24.80 ms**
32	0.5876	476.6	74.65	43.77 ms
64	**0.5865**	493.9	73.99	105.15 ms
128	0.5913	515.5	73.82	209.78 ms
WorldMem, 60 seconds				
16	**0.525**	**1041.75**	**74.28**	**3.52 ms**
32	0.534	1116.92	74.05	4.54 ms
64	0.545	1128.46	73.11	7.63 ms
128	0.577	1601.81	71.80	12.85 ms

Retention versus retrieval failure. We further decompose memory failure into two retrospective diagnostics: *retention gap*, measuring useful evidence lost through eviction, and *selection gap*, measuring the retriever's shortfall relative to the best evidence still retained. Figure 3c shows full retention removes eviction loss but still exhibits the largest selection gap. In contrast, bounded policies trade a small amount of retention loss for a substantially smaller selection gap. This suggests that a compact archive can make useful evidence easier to retrieve, rather than simply reducing storage.

5 DISCUSSION AND CONCLUSION

We introduced KEEPSAKE, a training-free controller that maintains fixed-capacity spatial memory by reassessing geometric and visual substitutability as new views arrive. Across MemCam and WorldMem, KEEPSAKE improves FVD and LPIPS over complete retention without changing the generator or retrieval rule. On fifteen matched 180-second MemCam trajectories, it reduces FVD by 35.1% while retaining 32 rather than 5,397 observations. Our diagnostics further distinguish evidence lost through eviction from retained evidence left unselected by retrieval, highlighting archive maintenance as a complement to retrieval. Together, these results demonstrate that selective retention can improve generation quality, not merely reduce storage.

AI USE STATEMENT

In this work, we have used LLMs to retouch paragraphs. We also used AI tools to revise formatting issues for figures and tables.

ETHICS STATEMENT

We do not include any human-related data or anything harmful that might raise ethical issues. Our scope is memory budgeting for long horizon interactive video generation.

REPRODUCIBILITY STATEMENT

We include all our memory policy code in the supplementary materials. We also provide full implementation details in the appendix. Our method is reproducible and testable with other papers due to its plug-and-play nature as well.

REFERENCES

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

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320. IEEE, 2024.

Xinhang Gao, Junlin Guan, Shuhan Luo, Wenzhuo Li, Guanghuan Tan, and Jiacheng Wang. Memcam: Memory-augmented camera control for consistent video generation. *arXiv preprint arXiv:2603.26193*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

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

Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

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

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

Weijie Kong, Q Tian, Z Zhang, R Min, Z Dai, J Zhou, J Xiong, X Li, B Wu, J Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models, 2025. *URL https://arxiv.org/abs/2412.03603*, 2(3), 2024.

Kunyang Li, Mubarak Shah, and Yuzhang Shang. Packcache: A training-free acceleration method for unified autoregressive video generation via compact kv-cache. *arXiv preprint arXiv:2601.04359*, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025.

T Li et al. Realcam-i2v: real-world image-to-video generation with interactive complex camera control (2025). *arXiv preprint arXiv:2502.10059*.

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

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Shuwei Shi, Zhen Li, Muyao Niu, Chuanhao Li, Bo Zheng, Kaipeng Zhang, and Yinqiang Zheng. Surprise forcing: What to remember, when to skip in long video generation. *arXiv preprint arXiv:2607.18436*, 2026.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Zun Wang, Han Lin, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. Anchorweave: World-consistent video generation with retrieved local spatial memories. In *European Conference on Computer Vision*, pp. 203–224. Springer, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026.

Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *International Conference on Learning Representations*, volume 2025, pp. 102934–102948, 2025.

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.

Tian-Xing Xu, Zi-Xuan Wang, Guangyuan Wang, Li Hu, Zhongyi Zhang, Peng Zhang, Bang Zhang, and Song-Hai Zhang. Ucm: Unified modeling of camera control and memory with time-aware positional encoding warping for world models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

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

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, Animesh Garg, et al. Mosaicmem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. Worldstereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories. *arXiv preprint arXiv:2603.02049*, 2026.

J Zhao, F Wei, Z Liu, H Zhang, C Xu, and Y Lu. Spatia: Video generation with updatable spatial memory (2025).

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

A Implementation and Evaluation Details

A.1 Evaluation Suites and Matching

Extended rollouts. We evaluate MemCam (Gao et al., 2026) on Context-as-Memory (Yu et al., 2025) and WorldMem (Xiao et al., 2026) on its Minecraft world-simulation dataset. The primary evaluation contains fifteen matched 60-second videos per system. A separate MemCam evaluation contains fifteen matched 180-second trajectories. Results are reported separately by system and duration. The MemCam suite also informed policy development.

Within each extended-rollout suite, retention policies share the initial inputs, prescribed control sequences, generator checkpoint, generation settings, and random seeds. Only the persistent-memory update changes. Each policy conditions generation on its own retained archive, allowing retention decisions to affect subsequent outputs. The generator and retrieval rule remain unchanged.

Native camera round trips. We additionally evaluate MemCam + KEEPSAKE using MemCam's camera-round-trip protocol on Context-as-Memory. The two settings contain 90° round trips of 153 frames and 360° round trips of 609 frames. KEEPSAKE uses $B = 32$ in both settings. This benchmark is reported separately from the extended-rollout suites.

A.2 Host Configurations and Memory Interfaces

Each MemCam and WorldMem generation job was allocated one NVIDIA H100 PCIe GPU, 16 CPUs, and 64 GiB of host memory. The `memcam` environment uses Python 3.10 and PyTorch 2.5.1 (CUDA 12.4 build), with the `cuda/cuda-12.1.0` module loaded.

MemCam uses Wan2.1 1.3B (Wan et al., 2025) with 76-frame chunks, 50 denoising steps, and ℓ_2-normalized DINOv2-Base (Oquab et al., 2023) retention descriptors. WorldMem uses its released Minecraft checkpoints at 128×128 resolution with 20 DDIM steps ($\eta = 0$); retention descriptors are VAE latents pooled to 4×4, flattened, and ℓ_2-normalized. KEEPSAKE preserves native memory representations and retrieval rules: MemCam retrieves one frame per query for 76 context slots per chunk, while WorldMem retrieves eight observations per query.

A.3 Metrics and Reporting

Generation quality. LPIPS (Zhang et al., 2018) compares generated frames with ground truth at corresponding trajectory indices, while FVD (Unterthiner et al., 2018) measures video-level distributional quality. Lower values are better for both. The native round-trip benchmark additionally reports PSNR and SSIM, for which higher values are better. Within each extended-rollout suite, policies are evaluated on the same video cohort with the same metric configuration.

We evaluate standard VBench (Huang et al., 2024) dimensions: subject consistency, background consistency, aesthetic quality, and imaging quality. Scores are multiplied by 100. Together, these dimensions describe perceptual quality and temporal behavior alongside the reference-based fidelity metrics.

Resource accounting. The memory budget counts retained observations, rather than the context slots filled for an individual generation step. Stored-item counts describe the archive; they are distinct from GPU memory use, host-memory allocation, and storage of generated outputs. Retrieval latency and total generation time are likewise separate quantities.

B Retention Policies and Update Details

B.1 Retention Baselines

The primary comparisons use complete retention, FIFO, K-center, Marginal Coverage Eviction (MCE), and KEEPSAKE. All bounded policies use $B = 32$ in the main comparison. They operate through the host's persistent-memory interface without replacing its retrieval rule.

Algorithm 2 Archive-Based K-Center Retention

Require: Candidates P, historical archive H, budget B
Require: Protected frames F, distance function d
Ensure: Retained bank S, with $|S| \leq B$

if $|P| \leq B$ **then**
  return P
end if
if $H = \emptyset$ **then**
  $H \leftarrow P$
end if
$S \leftarrow F$
if $S = \emptyset$ **then**
  $p_0 \leftarrow \arg\min_{p \in P} \frac{1}{|H|} \sum_{h \in H} d(h, p)$
  $S \leftarrow \{p_0\}$
end if
while $|S| < B$ **do**
  $h^\star \leftarrow \arg\max_{h \in H} \min_{s \in S} d(h, s)$
  $p^\star \leftarrow \arg\min_{p \in P \setminus S} d(h^\star, p)$
  $S \leftarrow S \cup \{p^\star\}$
end while
return S

Complete retention. All observations admitted by the host remain available for retrieval. The archive grows as new observations arrive, while the number of retrieved context slots follows the original host configuration.

FIFO. When the archive exceeds the budget, FIFO discards the oldest observations eligible for eviction until the capacity constraint is satisfied. Retention is therefore driven by recency rather than geometric or visual substitutability.

K-center. K-center uses geometry and appearance to select representative observations, targeting the distance between the least-covered historical view and the retained archive shown in Algorithm 2. Its objective emphasizes worst-case coverage rather than recency or the number of substitutes for an individual observation.

Marginal Coverage Eviction. MCE uses geometry and appearance to measure collective coverage of representative historical queries. It repeatedly removes the observation whose deletion causes the smallest decrease in coverage shown in Algorithm 3. Unlike KEEPSAKE's fixed within-update ranking, MCE bases each removal on the marginal coverage contribution to the retained set.

B.2 Memory Updates and Protection

At update t, the retained bank M_{t-1} and incoming observations N_t form the candidate set $V_t = M_{t-1} \cup N_t$. If $|V_t| \leq B$, KEEPSAKE retains all candidates. Otherwise, it removes $|V_t| - B$ eviction-eligible observations according to the priorities in Eq. 7.

Protected observations are excluded from eviction but count toward B. They are not additional memory outside the budget. Protection must therefore leave enough eligible observations to satisfy the capacity constraint. The resulting archive remains a subset of V_t in the host's native representation.

B.3 Numerical Conventions and Default Parameters

Camera poses are represented as camera-to-world rotations and translations. Translation distances are normalized by the median nonzero pairwise translation distance in the current candidate set, with a floor of $\epsilon = 10^{-6}$. When all positions coincide, the translation scale is set to one. Visual

Algorithm 3 Marginal Coverage Eviction (MCE)

Require: Candidates P, budget B, protected frames F
Require: Queries Q, weights w_q, similarities $K(q, m) \in [0, 1]$
Ensure: Retained bank S, with $|S| \leq B$
1: Define coverage:

$$C(S) = \sum_{q \in Q} w_q \left(1 - \prod_{m \in S} (1 - K(q, m))\right).$$

2: $S \leftarrow P$
3: **while** $|S| > B$ **do**
4: $\quad m^\star \leftarrow \arg\min_{m \in S \setminus F} \big[C(S) - C(S \setminus \{m\})\big]$
5: $\quad S \leftarrow S \setminus \{m^\star\}$
6: **end while**
7: **return** S

descriptors are ℓ_2-normalized, and negative cosine similarities are clipped to zero when constructing the pose-appearance affinities.

Self-comparisons are excluded from both the substitute count and maximum similarity. An edge is included when its affinity is at least T, whereas maximum similarity is computed over all other candidates, including affinities below T. Priorities are computed once over V_t and remain fixed throughout that update's evictions. Equal-priority ties are broken by earlier insertion. The next update recomputes priorities using the new candidate set.

Table 4 lists the shared parameter settings. The budget sweep varies $B \in \{16, 32, 64, 128\}$ while holding the other retention parameters fixed.

Table 4: **Default retention settings.** The same values are used for MemCam and WorldMem.

Parameter	Role	Value
B	Retained-observation budget	32
α	Pose–appearance weighting	0.65
T	Strong-substitute threshold	0.65
τ	Degree saturation threshold	3
β	Inverse-degree weight	0.5
λ	Closest-substitute weight	0.25

C Additional Numerical Results and Resource Scaling

C.1 Long-Horizon Generation

Table 5 provides the numerical values underlying Fig. 3a. On the fifteen matched 180-second MemCam trajectories, KEEPSAKE improves both LPIPS and FVD over complete retention and FIFO while maintaining $B = 32$.

Table 5: **Long-horizon MemCam results.** Fifteen matched 180-second trajectories; bounded policies use $B = 32$.

Policy	Retained items	LPIPS ↓	FVD ↓
Complete retention	5,397	0.5980	734.2
FIFO	32	0.6514	677.3
KEEPSAKE	32	**0.5876**	**476.6**

Table 6: End-to-end generation time per completed video. Comparisons are made within each system.

System	Rollout	Unbounded	KEEPSAKE B32	Reduction	Speedup
MemCam	60 s	106.46 min	**65.46 min**	**41.00 min (38.5%)**	**1.626**×
MemCam	180 s	587.99 min	**193.14 min**	**394.85 min (67.2%)**	**3.044**×
WorldMem	60 s	12.30 min	**11.85 min**	**0.45 min (3.7%)**	**1.038**×

C.2 BUDGET SENSITIVITY

The budget sweep in Table 3 evaluates MemCam at 180 seconds and WorldMem at 60 seconds. In MemCam, the lowest reported FVD occurs at $B = 16$, whereas the lowest LPIPS occurs at $B = 64$. In WorldMem, both metrics are lowest at $B = 16$ among the tested budgets. Retrieval time increases with capacity in both systems. These results characterize the operating tradeoff of the retention policy: increasing the budget does not consistently improve generation quality.

C.3 ARCHIVE AND RETRIEVAL SCALING

Let s be the maximum number of observations added per chunk. Since the retained bank contains at most B observations, a KEEPSAKE update processes at most $B + s$ candidates. For fixed descriptor dimension, direct computation of all pairwise affinities requires $O((B + s)^2)$ work and temporary storage. Sorting the priorities adds $O((B + s) \log(B + s))$ work. The persistent archive requires $O(B)$ storage for fixed observation sizes.

With a fixed number of retrieval queries per chunk and fixed initial history, complete retention requires $O(ns)$ archive items and $O(n^2 s)$ cumulative exhaustive candidate comparisons over n chunks. Bounded retention uses at most B archived observations and $O(nB)$ cumulative comparisons. These bounds describe the retained archive and candidate-scoring work, not the generator's weights, activations, or output buffers.

For the 180-second MemCam evaluation, the mean unbounded candidate count grows from 671.9 in the first 45 seconds to 4,689.5 in the last 45 seconds. KEEPSAKE caps the retained bank at 32 after each update without changing the number of retrieved context slots.

D RETENTION AND RETRIEVAL DIAGNOSTICS

D.1 GENERATION EVALUATIONS AND COMMON-SOURCE DIAGNOSTICS

In the generation evaluations, each policy produces its own video while updating and querying its own archive. The resulting scores therefore reflect the interaction between retention and subsequent generation. The common-source diagnostics instead load policy-selected historical images from the same complete-retention rollout. This separates differences in the selected historical evidence from differences between independently generated videos.

Ground-truth references are used retrospectively to evaluate the selected evidence. They are not inputs to the online retention priority, which uses camera poses and appearance descriptors from the observations currently available to the controller.

D.2 VIEW RELEVANCE AND HISTORICAL-IMAGE FIDELITY

Figure 3b distinguishes view mismatch from memory corruption. The former concerns agreement with the queried viewpoint, whereas the latter concerns the fidelity of the selected historical content. Both are measured using DINOv2 cosine distance in the fifteen-trajectory, 60-second common-source comparison. Table 7 gives the means and 95% trajectory-bootstrap confidence intervals shown in the figure.

KEEPSAKE has higher mean view mismatch but lower mean memory corruption than complete retention in this comparison. The two quantities therefore need to be distinguished: closer agreement with a query viewpoint does not necessarily coincide with higher historical-image fidelity.

Table 7: **Common-source memory diagnostics.** Fifteen 60-second trajectories; bounded policies use $B = 32$. Entries are means with 95% trajectory-bootstrap confidence intervals. Lower distances are better.

Policy	View mismatch	Memory corruption
Complete retention	0.1256 [0.1041, 0.1479]	0.4871 [0.4012, 0.5716]
FIFO	0.1557 [0.1336, 0.1777]	0.5022 [0.4195, 0.5835]
KEEPSAKE	0.3254 [0.2579, 0.3964]	0.4079 [0.3159, 0.4989]

D.3 RETENTION GAP AND SELECTION GAP

The retention-selection diagnostic separates two sources of shortfall relative to useful historical evidence. Retention gap measures evidence lost through eviction by comparing the best available full-history evidence with the best evidence still retained. Selection gap measures the retriever's shortfall relative to the best retained evidence. The comparison uses target ground-truth features retrospectively; neither gap is an online retention objective.

Table 8 records the rounded coordinates in Fig. 3c. Complete retention has zero retention gap by construction and a selection gap of 0.2188. At each displayed budget, KEEPSAKE has a smaller retention gap than FIFO, while FIFO has a smaller selection gap. Increasing capacity reduces retention gap but increases selection gap along both plotted policy curves. The diagnostic thus separates preserving evidence from having the existing retriever select it.

Table 8: **Retention-selection diagnostic at 180 seconds.** Rounded coordinates from Fig. 3c. Complete retention has retention gap 0 and selection gap 0.2188.

	FIFO		KEEPSAKE	
Budget B	Retention gap	Selection gap	Retention gap	Selection gap
16	0.1872	0.0448	0.0669	0.1037
32	0.1668	0.0646	0.0426	0.1470
64	0.1429	0.0867	0.0366	0.1596
128	0.1203	0.1084	0.0298	0.1783

D.4 SENSITIVITY OF THE HISTORICAL-FIDELITY COMPARISON

For the PSNR and SSIM analysis, selected historical images are loaded from the same complete-retention rollout and compared with ground truth at their own historical indices. This evaluates the fidelity of stored content rather than agreement with the current query viewpoint. Across fifteen trajectories, the aggregate KEEPSAKE-minus-complete-retention differences are $+4.629$ dB PSNR and $+0.1512$ SSIM.

Table 9 shows how these differences change under two successive restrictions. First, we exclude selection disagreements in which KEEPSAKE chooses the initial conditioning image. We then additionally require both selected observations to have field-of-view overlap of at least 0.80.

Table 9: **Sensitivity of historical-image fidelity.** Differences are KEEPSAKE minus complete retention on common-source images, evaluated against ground truth at their historical indices. The restrictions are cumulative.

Comparison subset	ΔPSNR (dB)	ΔSSIM
Aggregate comparison	+4.629	+0.1512
Exclude initial-image selection disagreements	+1.060	+0.0449
Additionally require FOV overlap ≥ 0.80 for both selections	+0.103	+0.0098

Both confidence intervals for the final restricted comparison include zero. The aggregate advantage therefore depends on the selection distribution and does not establish a general ability to identify higher-fidelity memories at matched viewpoints.

E QUALITATIVE COMPARISONS

We provide two sets of qualitative comparisons. Figure 5 presents additional camera-revisit examples comparing unbounded retention, FIFO, and KEEPSAKE. Figure 6 presents selected common-source queries for which unbounded retention and KEEPSAKE retrieve different historical observations.

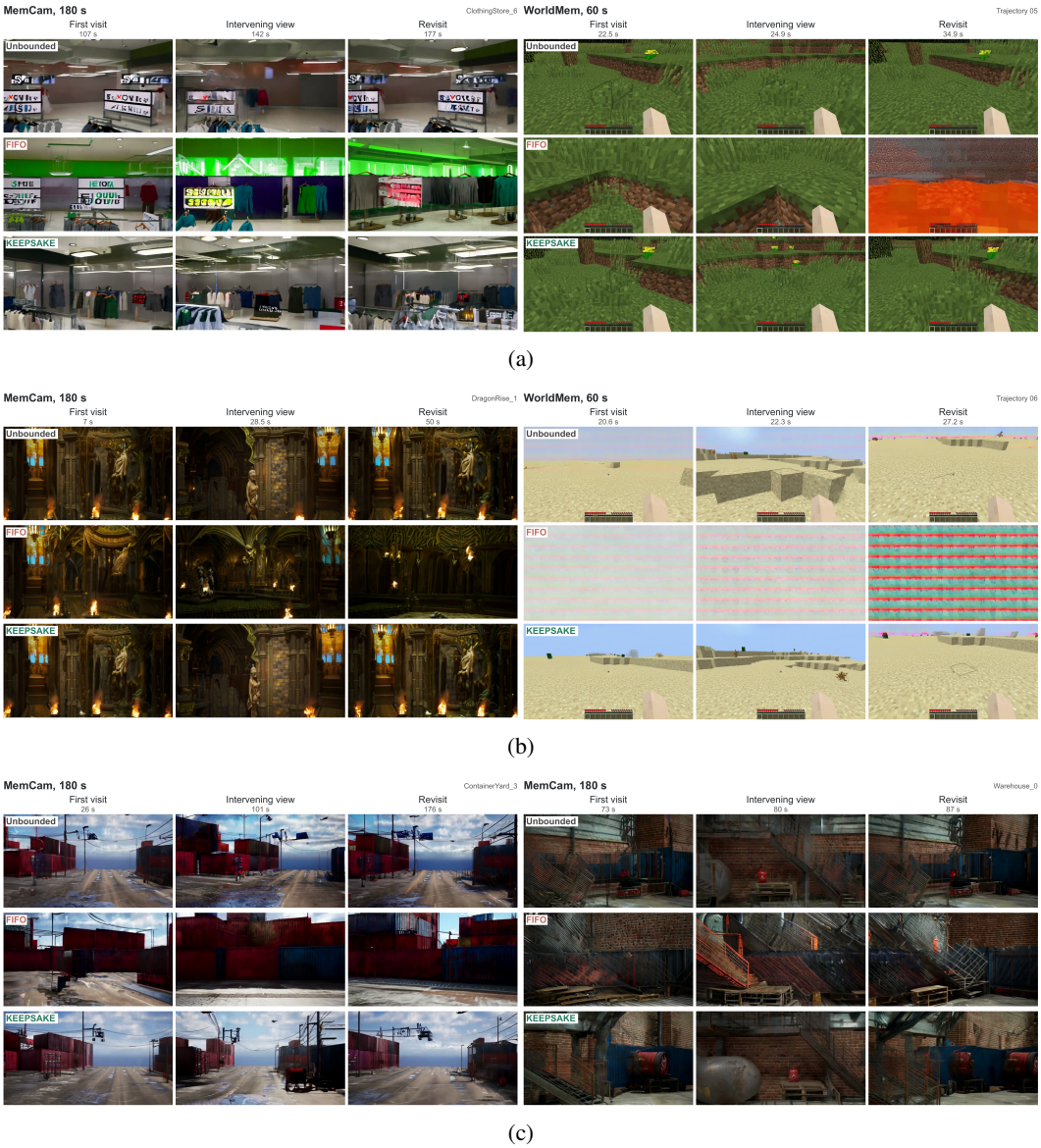


Figure 5: **Additional camera-revisit examples.**

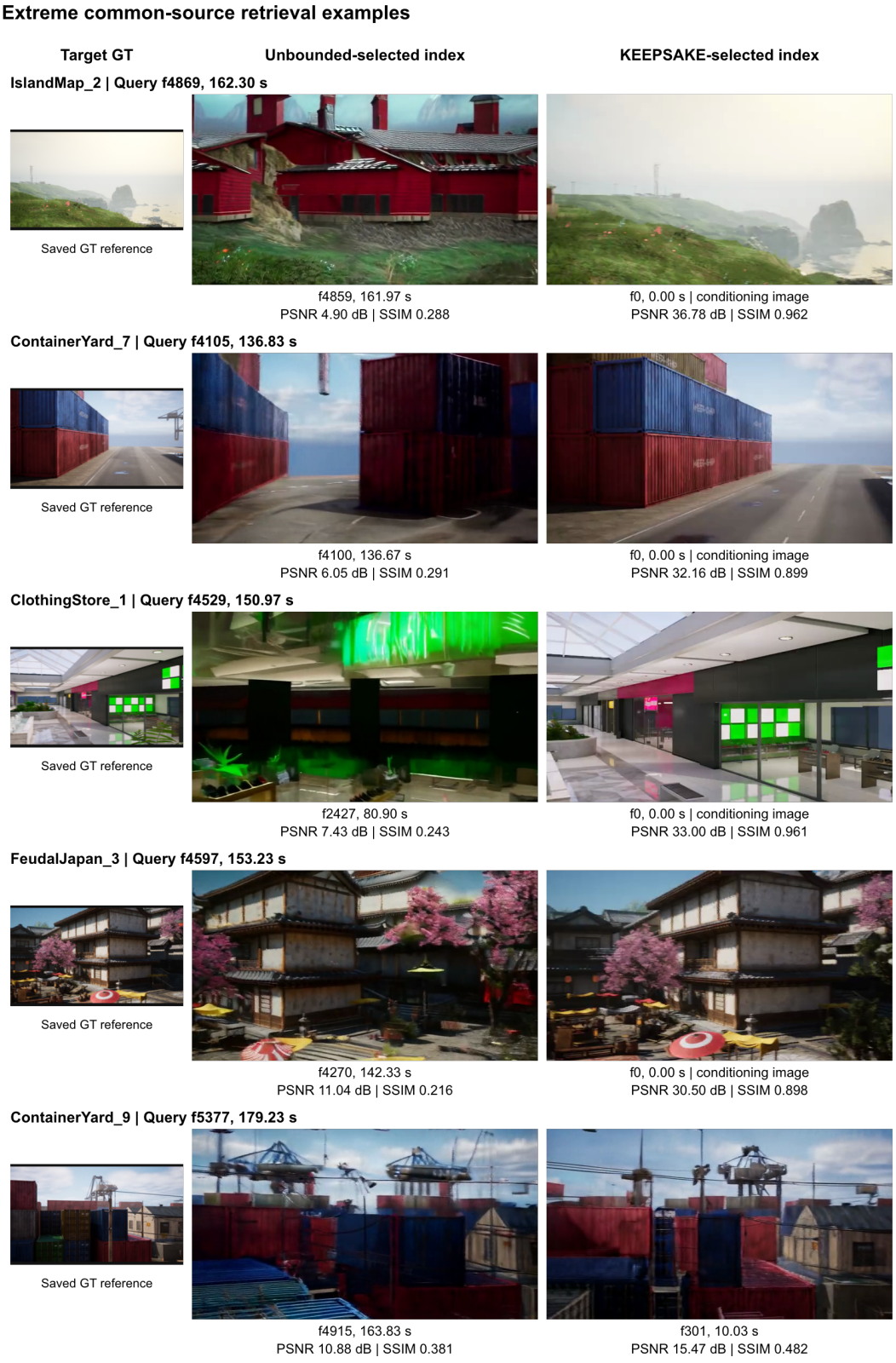


Figure 6: Selected common-source retrieval examples. For identical queries, unbounded retention and KEEPSAKE can retrieve substantially different historical observations.