CONSISTKV: LONG-TERM KV MEMORY FOR CONSISTENT INTERACTIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world models can generate controllable visual environments in real time, yet they often fail to preserve previously observed scenes over long interactions. We identify this failure as *revisit collapse*: once historical observations leave the bounded KV context, the model loses direct access to their visual states and may reconstruct different content when revisiting the same location. We propose *ConsistKV*, a lightweight external KV memory for frozen interactive world models that maintains compressed distant history without extending the active attention window. For retrieval, our *Action–Content Encoder (ACE)* jointly models visual content and action-derived relative geometry, combining appearance correspondence with complementary geometric localization. We further introduce a training-free *Value-Absorbing KV Compression* strategy that reduces historical KV storage while preserving information from discarded positions through shared self-Gram-guided selection and value absorption. Experiments on Matrix-Game 2.0 and ABot-World show improved revisit consistency over sliding-window and fixed-budget memory baselines, with clearer gains over longer interaction horizons.

1 INTRODUCTION

Recent advances in generative world models have enabled real-time, interactive environments that synthesize future observations in response to user actions (Bruce et al., 2024; He et al., 2025; Team et al., 2026; Jiang et al., 2026). However, sustaining a persistent world over long interactions remains challenging (Ye et al., 2026; Yi et al., 2026; Wu et al., 2026; Gu et al., 2026; Garcin et al., 2026). Such persistence is essential for meaningful long-horizon interaction: after exploring an environment and returning to a previously visited location, users naturally expect its layout, landmarks, and surrounding structure to remain consistent with what was previously observed. Yet current models often fail to preserve such previously established scene states, causing objects, layouts, or other scene content to change upon revisit; we refer to this failure as *revisit collapse*. One major cause is the bounded KV context required for efficient autoregressive generation: older KV states are evicted as new scenes are generated, removing direct access to previously observed regions. Retaining the complete KV history can mitigate this forgetting, but its storage and attention costs grow with rollout length, conflicting with real-time interactive generation (Yi et al., 2026; Wu et al., 2026).

A promising alternative is to preserve distant history outside the active context and retrieve only the states relevant to the current generation. This strategy has been explored in long-form video generation, where compact history and content-based retrieval are used to extend temporal consistency beyond the local window (Yang et al., 2025; Hu et al., 2026). Interactive world models additionally provide action-derived geometric cues that can help localize relevant past states (Yu et al., 2025; Hong et al., 2025; Wang et al., 2026; Yi et al., 2026; Chen et al., 2026; Xu et al., 2026). These cues provide complementary information: visual similarity provides a direct cue for appearance correspondence but can be ambiguous across self-similar locations, while action-derived geometry provides localization without directly encoding the visual correspondence between historical and current observations. Moreover, retrieval-based memory must retain a sufficiently long history while keeping the storage cost of historical states manageable. Simply pruning historical KV entries reduces storage, but may discard information needed when those states are revisited. This motivates both retrieval that combines complementary cues and effective KV compression that retains useful information from discarded entries.

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 address these challenges, we propose *ConsistKV*, a lightweight long-term KV memory framework that combines action–content retrieval, Value-Absorbing KV Compression, and fixed-budget memory delivery for frozen interactive world models. ConsistKV maintains compact historical KV states outside the native context and selectively recalls relevant memory during long-horizon interaction. For retrieval, we introduce an *Action–Content Encoder (ACE)* that jointly exploits visual content and action-derived relative geometry, combining appearance correspondence with complementary geometric localization. ACE is trained from automatically constructed revisit correspondences without manual annotation.

For historical KV compression, *ConsistKV* introduces a training-free *Value-Absorbing KV Compression* strategy that compactly preserves historical KV states by sharing a self-Gram structure between representative-position selection and discarded-value absorption. Retrieved states are then relocated into the active context through backbone-compatible *memory delivery*, allowing long-range history to be recalled under a fixed active attention budget. Consequently, only the lightweight ACE module is learned, while KV compression, memory delivery, and the world-model backbone require no additional training. Experiments on Matrix-Game 2.0 (He et al., 2025) and ABot-World (Jiang et al., 2026) demonstrate improved revisit consistency over sliding-window and existing memory baselines, with particularly clear gains under long-horizon interactions.

Our contributions are summarized as follows:

- We propose *ConsistKV*, a lightweight external KV memory that gives frozen interactive world models access to distant historical states while preserving their native active attention budget, thereby improving consistency across long-horizon revisits.
- We introduce *ACE*, which jointly models visual content and action-derived relative geometry to combine appearance correspondence with geometric localization for historical retrieval, and learns from automatically constructed revisit correspondences.
- We introduce a training-free *Value-Absorbing KV Compression* strategy that couples representative-position selection and discarded-value absorption through a shared self-Gram structure.

2 Related Work

2.1 Interactive World Models

Interactive world models generate controllable visual environments by predicting future observations conditioned on user actions (Bruce et al., 2024; Valevski et al., 2025; He et al., 2025; Li et al., 2025; Team et al., 2026; Jiang et al., 2026). Recent advances have substantially improved generation quality, action controllability, and real-time interaction, while long-term consistency remains challenging (Ye et al., 2026). To maintain persistent world states, recent methods augment the local context with historical frame or latent memories (Xiao et al., 2026; Yu et al., 2025; Sun et al., 2025), compact long-term representations (Hong et al., 2025; Wang et al., 2026; Chen et al., 2026; Peng et al., 2026), or explicit spatial memories (Yu et al., 2026; Garcin et al., 2026). These approaches demonstrate the importance of retaining historical information for persistent interactive generation.

2.2 Long-Horizon Memory and Retrieval

Long-form video generation has explored efficient access to distant history beyond the local temporal context through compact temporal memories (Yang et al., 2025), sparse routing or attention to historical states (Cai et al., 2026; Zhao et al., 2026a), and content-based retrieval (Hu et al., 2026). Interactive world models further provide camera, pose, and action signals that can guide historical localization, enabling retrieval based on view overlap or accumulated geometry (Yu et al., 2025; Yi et al., 2026; Chen et al., 2026), as well as selective access to long-term context through sparse memory mechanisms (Xu et al., 2026). Recent approaches also maintain compressed KV memories that remain addressable over long rollouts under positional constraints (Wu et al., 2026). Rather than relying on a single retrieval signal, our method jointly models visual content for appearance correspondence and action-derived relative geometry for localization, combining complementary cues for historical retrieval.

2.3 KV Cache Compression

KV cache compression has been widely studied to reduce the memory and computational cost of autoregressive inference. Existing approaches retain informative tokens based on attention statistics (Zhang et al., 2023; Li et al., 2024), adaptively allocate cache budgets across layers (Cai et al., 2024), or exploit temporal redundancy in video KV states (Samuel et al., 2026; Ji et al., 2026; Zhao et al., 2026b). Recent world-model methods further compress historical KV states to extend the accessible interaction history under a fixed attention budget (Yi et al., 2026). Beyond eviction, prior work has explored preserving discarded information through token or KV merging (Bolya et al., 2022; Zhang et al., 2024; Wan et al., 2024; Wang et al., 2024; Yuan et al., 2025), or through an auxiliary recurrent cache (Dong et al., 2024). Our compression instead uses a shared self-Gram structure to couple representative-position selection with soft redistribution of discarded values, while leaving the retained keys unchanged to preserve their addresses for subsequent memory delivery in the interactive world model.

3 Preliminaries

Interactive World Models. Interactive world models (He et al., 2025; Team et al., 2026; Jiang et al., 2026) generate a video chunk by chunk, where each chunk contains a few frames (e.g., 3) and is conditioned on user actions. Let $\mathbf{x}_0$ denote the initial latent chunk and $\mathbf{x}_{1:T} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ the subsequently generated latent chunks, each containing a few latent frames (e.g., three in our setting). The action input $\mathbf{a}_t$ is applied after observing $\mathbf{x}_t$ to generate $\mathbf{x}_{t+1}$. Given the initial chunk and action sequence, the model factorizes the generation distribution as

$$p_\Theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0, \mathbf{a}_{0:T-1}) = \prod_{t=1}^{T} p_\Theta(\mathbf{x}_t \mid \mathbf{x}_{<t}, \mathbf{a}_{<t}), \tag{1}$$

where $\mathbf{x}_{<t} = (\mathbf{x}_0, \ldots, \mathbf{x}_{t-1})$ and $\mathbf{a}_{<t} = (\mathbf{a}_0, \ldots, \mathbf{a}_{t-1})$ denote the preceding chunks and actions, respectively, and p_Θ denotes the conditional distribution modeled by the pretrained world model with parameters Θ.

KV-Cached Autoregressive Generation. To avoid re-encoding the history at every step, autoregressive world models cache the self-attention key–value (KV) states of previously generated chunks. We denote the KV state of chunk $\mathbf{x}_t$ by $\mathbf{B}_t = (\mathbf{K}_t, \mathbf{V}_t)$. With *full KV caching*, every past KV state is kept, so $\mathbf{x}_t$ can attend to the entire history; however, both the attention cost per chunk and the cache memory grow linearly with the rollout length, which quickly becomes prohibitive for long-horizon interaction. Recent interactive world models therefore adopt *sliding-window* inference (He et al., 2025; Jiang et al., 2026), which keeps only the KV states of the latest L chunks. The KV cache available when generating $\mathbf{x}_t$ is:

$$\mathcal{C}_t = [\mathbf{B}_{t-L}, \ldots, \mathbf{B}_{t-1}], \tag{2}$$

where entries with negative indices are omitted at the beginning of the rollout. This bounds the inference cost, but at a price: once a chunk falls more than L steps behind, its KV state is evicted and can never influence generation again. Over a long horizon, the model thus loses all memory of earlier content; when the user returns to a previously visited place, the scene is synthesized anew and no longer matches what was generated before. We refer to this failure as *Revisit Collapse*.

4 Method

We propose *ConsistKV*, an external KV memory that gives frozen interactive world models access to their distant past. ConsistKV stores compressed *historical KV states* outside the active cache and selectively retrieves those relevant to the current generation. The retrieved states replace part of the sliding-window context, allowing the backbone to reuse distant observations while maintaining a fixed active attention budget. This access to historical information improves consistency across revisits and mitigates *revisit collapse*.

As illustrated in Figure 1, ConsistKV consists of three components. *Action–Content Retrieval* (section 4.1) selects relevant chunks from the historical KV storage by jointly considering latent visual

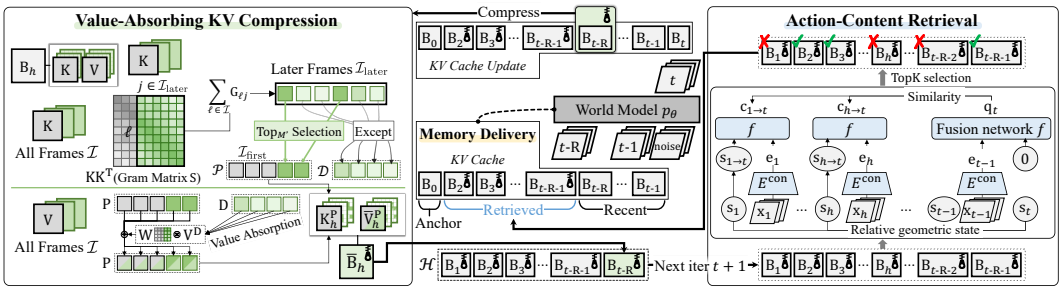


Figure 1: Overall pipeline of ConsistKV. *Action–Content Retrieval* (section 4.1) selects relevant entries from the historical storage using visual content and action-derived geometry, whose KV states are stored in compressed form by *Value-Absorbing KV Compression* (section 4.2). *Memory Delivery* (section 4.3) then places the retrieved compressed states into the active cache together with the persistent anchor $\mathbf{B}_0$ and recent native states under the original attention budget.

content and action-derived relative geometry. The retrieval module is the only learned component: it is trained offline before inference using automatically constructed revisit correspondences and is then kept fixed for retrieval during generation. At inference time, the fixed retrieval module operates together with two training-free components. *Value-Absorbing KV Compression* (section 4.2) uses a shared self-Gram structure to select representative positions and absorb discarded values while keeping the retained keys unchanged. Finally, *Memory Delivery* (section 4.3) relocates the retrieved KV states into the active cache while preserving the backbone's original attention budget.

4.1 Action–Content Retrieval

Retrieving a useful historical state requires both locating the relevant past region and identifying a visually corresponding observation. Visual content provides a direct cue for appearance correspondence, but can be ambiguous across different locations with similar appearance. Action-derived geometry provides a complementary localization cue, but does not directly capture visual correspondence between the historical observation and the current generated context. We therefore introduce an *Action–Content Encoder* (ACE) that jointly embeds latent visual content and action-derived relative geometry to score each entry in the historical KV storage.

Historical KV Storage. Just as the KV cache $\mathcal{C}_t$ in equation 2 is the context that the backbone attends to, we define the *historical KV storage* $\mathcal{H}_t$ as the pool that ACE scores before generating $\mathbf{x}_t$:

$$\mathcal{H}_t = [\mathbf{B}_1, \ldots, \mathbf{B}_{t-R-1}], \tag{3}$$

where the latest R chunks are excluded because they remain in the active cache in their native form, and the initial chunk is excluded because $\mathbf{B}_0$ is always kept in the active cache as a persistent anchor. The storage is never attended to directly: only the K entries selected by ACE enter the active cache, so its size does not add to the backbone's attention cost. For brevity, we write $h \in \{1, ..., t-R-1\}$ to index the entries of the storage.

Action–Content Encoding. ACE operates directly on the latent chunks $\mathbf{x}_h$ produced by the frozen world model. A lightweight content encoder E^{con} maps each chunk to a compact descriptor $\mathbf{e}_h = E^{\text{con}}(\mathbf{x}_h)$; it is trained from scratch together with the rest of ACE and requires no external pretrained visual encoder.

To provide geometric context, we maintain a *geometric state* $\mathbf{s}_t = (\gamma_t, p_t^x, p_t^z, \beta_t)$ by integrating the actions $\mathbf{a}_{<t}$, where γ_t and β_t denote yaw and pitch and (p_t^x, p_t^z) the accumulated planar position. Since $\mathbf{a}_{<t}$ is known before $\mathbf{x}_t$ is generated, the destination state $\mathbf{s}_t$ is available at retrieval time. We express the relative geometric state $\mathbf{s}_{h\to t}$ of each h relative to this destination t:

$$\mathbf{s}_{h\to t} = \begin{bmatrix} \mathrm{wrap}(\gamma_h - \gamma_t) \\ \Delta p^x \sin\gamma_t + \Delta p^z \cos\gamma_t \\ \Delta p^x \cos\gamma_t - \Delta p^z \sin\gamma_t \\ \beta_h - \beta_t \end{bmatrix}, \qquad (\Delta p^x, \Delta p^z) = (p_h^x - p_t^x,\ p_h^z - p_t^z), \tag{4}$$

where $\mathrm{wrap}(\cdot)$ maps an angle to $[-\pi, \pi)$. The four entries of $\mathbf{s}_{h\to t}$ are the relative yaw, forward displacement, lateral displacement, and relative pitch in the coordinate frame of the destination.

A lightweight fusion network $f(\cdot,\cdot)$ then maps a content descriptor $\mathbf{e}$ and a relative geometric state $\mathbf{s}$ to a joint, ℓ_2-normalized embedding $f(\mathbf{e}, \mathbf{s})$. Because the destination chunk $\mathbf{x}_t$ has not been generated yet, the latest chunk $\mathbf{x}_{t-1}$ serves as its visual proxy and is placed at the destination itself (*i.e.,* at zero relative geometry), whereas each historical entry is conditioned on its geometric state relative to the destination:

$$\mathbf{q}_t = f(\mathbf{e}_{t-1}, \mathbf{0}), \qquad \mathbf{c}_{h\to t} = f(\mathbf{e}_h, \mathbf{s}_{h\to t}). \tag{5}$$

The retrieval score is their cosine similarity, and ConsistKV retrieves the K highest-scoring entries:

$$r_{h\to t} = \mathbf{q}_t^\top \mathbf{c}_{h\to t}, \qquad \mathcal{R}_t = \mathrm{TopK}\left(\{r_{h\to t}\}_{h\in\mathcal{H}_t}\right), \tag{6}$$

where TopK returns the indices of the K largest elements.

Mirror-Retracing Supervision. Training ACE requires revisits with known correspondence (*i.e.,* a query together with the historical chunk it should retrieve), but we do not assume access to manually annotated revisit correspondences. We obtain them automatically through mirror-retracing trajectories. Given a forward rollout $\mathbf{x}_{1:n}$, we continue the rollout by issuing the inverse of each forward action in reverse order (*e.g.,* a backward step for a forward step, a right turn for a left turn), which yields an inverse rollout $\mathbf{x}_{n+1:2n}$. By construction, the accumulated geometric state of $\mathbf{x}_{n+k}$ matches that of the corresponding forward chunk $\mathbf{x}_{n-k}$ (*i.e.,* $\mathbf{s}_{n-k\to n+k} = \mathbf{0}$). For each inverse chunk $\mathbf{x}_t$, we therefore take the corresponding forward chunk with the matching accumulated geometric state as the positive h^+ and all other entries of $\mathcal{H}_t$ as negatives. We then score every entry of $\mathcal{H}_t$ exactly as at inference: the latent content descriptors and relative geometric states are combined by equation 5 and compared using equation 6. Since the resulting scores $r_{h\to t}$ depend on the chunks $\mathbf{x}$ and geometric states $\mathbf{s}$ only through E^{con} and f, both networks are trained jointly by pushing the score of the positive above those of the negatives with the contrastive loss:

$$\mathcal{L}_{\mathrm{ACE}} = -\log \frac{\exp(r_{h^+\to t}/\tau_c)}{\sum_{h\in\mathcal{H}_t}\exp(r_{h\to t}/\tau_c)}, \tag{7}$$

where τ_c is the contrastive temperature. This supervision is used only to train ACE offline; once trained, ACE remains fixed during inference and requires neither revisit correspondences nor further model updates.

4.2 Value-Absorbing KV Compression

As the rollout grows, retaining the native KV states of all historical chunks becomes memory-intensive. ConsistKV therefore compresses each historical KV state $\mathbf{B}_h$ into $\bar{\mathbf{B}}_h$ before committing it to historical storage. WorldKV (Yi et al., 2026) shows that temporal redundancy in video KV states is revealed by key-space similarity. Building on this observation, we compute a single self-Gram matrix and reuse it to couple two operations: selecting representative KV positions and absorbing discarded values into key-similar retained positions. Since each historical chunk is compressed independently, we omit the chunk index h below and restore it when referring to the compressed state committed to historical storage.

Each chunk comprises F latent frames with M (i.e., the number of KV tokens) KV positions per frame, so that $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{FM\times d}$ for an attention head of dimension d. Let $\mathcal{I} = \{1, \ldots, FM\}$ index all KV positions, split into the first frame $\mathcal{I}_{\mathrm{first}} = \{1, \ldots, M\}$ and the later frames $\mathcal{I}_{\mathrm{later}} = \{M+1, \ldots, FM\}$. The shared self-Gram matrix is

$$\mathbf{S} = \mathbf{K}\mathbf{K}^\top \in \mathbb{R}^{FM\times FM}. \tag{8}$$

Self-Gram KV Selection. The first latent frame is kept in full, and M' representative positions are selected from the later frames. Redundant visual content produces recurring structure in key space, so a position whose key is similar to many others is a good representative of them. We quantify this by row-normalizing the $\mathbf{S}$ and summing the similarity that each later-frame position receives:

$$\mathbf{G} = \mathrm{Softmax}\left(\frac{\mathbf{S}}{\sqrt{d}}\right), \qquad g_j = \sum_{\ell\in\mathcal{I}} \mathbf{G}_{\ell j}, \quad j \in \mathcal{I}_{\mathrm{later}}, \tag{9}$$

Algorithm 1 ConsistKV inference

Require: Frozen world model p_Θ; ACE (E^{con}, f) trained with equation 7; initial chunk $\mathbf{x}_0$ and actions $\mathbf{a}_{0:T-1}$; recent window R, retrieval size K.

Encode $\mathbf{x}_0$ to obtain the anchor KV state $\mathbf{B}_0$; set $\mathbf{s}_0 \leftarrow \mathbf{0}$, $\mathcal{H} \leftarrow \emptyset$
for $t = 1, \dots, T$ **do**
  Update the geometric state $\mathbf{s}_t$ by integrating $\mathbf{a}_{<t}$
  if $t > R + 1$ **then**
    // — Value-Absorbing KV Compression —
    $h \leftarrow t - R - 1$
    Select preserved KV positions $\mathcal{P}$ of $\mathbf{B}_h$ by equation 10
    Compute the absorbed values $\bar{\mathbf{V}}_h^{\text{P}}$ by equation 11
    $\bar{\mathbf{B}}_h \leftarrow (\mathbf{K}_h^{\text{P}}, \bar{\mathbf{V}}_h^{\text{P}})$; $\mathcal{H} \leftarrow \mathcal{H} \cup \{(\bar{\mathbf{B}}_h, \mathbf{e}_h, \mathbf{s}_h)\}$
  end if
  if $\mathcal{H} = \emptyset$ **then**
    Assign the native cache $\mathbf{B}_{0:t-1}$ to $\widetilde{\mathcal{C}}_t$
  else
    // — retrieve and assemble the active cache —
    Retrieve $\mathcal{R}_t \subset \mathcal{H}$ by equation 4–equation 6
    Assemble $\widetilde{\mathcal{C}}_t$ from $\mathbf{B}_0$, $\{\mathcal{A}_t(\bar{\mathbf{B}}_h)\}_{h \in \mathcal{R}_t}$, and $\mathbf{B}_{t-R:t-1}$ by equation 12
  end if
  Sample $\mathbf{x}_t \sim p_\Theta(\cdot \mid \widetilde{\mathcal{C}}_t, \mathbf{a}_{<t})$ by equation 1; cache $\mathbf{B}_t$ and $\mathbf{e}_t \leftarrow E^{\text{con}}(\mathbf{x}_t)$
end for
return $\mathbf{x}_{1:T}$

where the Softmax is applied row-wise. Equivalently, g_j is the total attention that position j receives when the keys of the chunk attend to themselves. The scores are summed over attention heads so that all heads share one set of preserved positions:

$$\mathcal{P} = \mathcal{I}_{\text{first}} \cup \text{Top}_{M'}\left(\{g_j\}_{j \in \mathcal{I}_{\text{later}}}\right), \tag{10}$$

where $\text{Top}_{M'}$ returns the indices of the M' largest scores; this reduces the chunk from FM to $M + M'$ KV positions.

Gram-Kernel Value Absorption. Selection decides which positions survive, but simply dropping the others would also discard the information carried by their values. We therefore reuse $\mathbf{S}$ to transfer each discarded value to the preserved positions with similar keys. Let $\mathcal{D} = \mathcal{I} \backslash \mathcal{P}$ denote the discarded positions, let the superscripts P and D denote the restriction of $\mathbf{K}$ and $\mathbf{V}$ to the rows in $\mathcal{P}$ and $\mathcal{D}$, and let $\mathbf{S}^{\text{D}\to\text{P}} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{P}|}$ be the submatrix of $\mathbf{S}$ with rows in $\mathcal{D}$ and columns in $\mathcal{P}$. The discarded values $\mathbf{V}^{\text{D}}$ are absorbed to $\mathbf{V}^{\text{P}}$ as follows:

$$\mathbf{W} = \text{Softmax}\left(\frac{\mathbf{S}^{\text{D}\to\text{P}}}{\tau_a}\right), \qquad \bar{\mathbf{V}}^{\text{P}} = \mathbf{V}^{\text{P}} + \alpha \mathbf{W}^\top \mathbf{V}^{\text{D}}, \tag{11}$$

where the temperature τ_a controls the sharpness of this distribution and α the strength of absorption. Each row of $\mathbf{W}$ distributes one discarded value over the preserved positions according to key similarity. Restoring the chunk index, the compressed state committed to historical storage is $\bar{\mathbf{B}}_h = (\mathbf{K}_h^{\text{P}}, \bar{\mathbf{V}}_h^{\text{P}})$, where the preserved keys remain unchanged and the discarded keys and values are removed after their value information has been transferred to the preserved positions.

4.3 Memory Delivery

Finally, Memory Delivery assembles the active cache that the backbone attends to when generating $\mathbf{x}_t$. It consists of the initial KV state $\mathbf{B}_0$ as a persistent anchor, the K compressed KV states $\{\bar{\mathbf{B}}_h\}_{h \in \mathcal{R}_t}$ retrieved from $\mathcal{H}_t$ ($\mathcal{R}_t$ from equation 6), and the native KV states of the latest R chunks, which preserve short-term continuity:

$$\widetilde{\mathcal{C}}_t = \big[\, \underbrace{\mathbf{B}_0}_{\text{anchor}} ; \underbrace{(\widetilde{\mathbf{B}}_h)_{h \in \mathcal{R}_t}}_{\text{retrieved}} ; \underbrace{\mathbf{B}_{t-R}, \dots, \mathbf{B}_{t-1}}_{\text{recent}} \,\big], \qquad \widetilde{\mathbf{B}}_h = \mathcal{A}_t(\bar{\mathbf{B}}_h). \tag{12}$$

Here, the retrieved chunks $\bar{\mathbf{B}}_h$ are sorted by their original index h, and $\mathcal{A}_t$ is a positional adaptation, required because the retrieved KV states are relocated from their original cache positions to the temporal slots vacated by the non-recent chunks they replace. Since the retrieved states replace the non-recent portion of the native cache $\mathcal{C}_t$ rather than being appended to it, $\widetilde{\mathcal{C}}_t$ stays within the backbone's original attention budget. The backbone-specific implementation of $\mathcal{A}_t$ depends on how each model stores and applies positional encoding; we provide these implementation details in Appendix A. We summarize the overall process of ConsistKV in Algorithm 1.

5 Experiments

5.1 Implementation Details

Backbones and Memory Configuration. We evaluate ConsistKV on two major backbones: Matrix-Game 2.0 (MG) (He et al., 2025) and ABot-World (ABot) (Jiang et al., 2026). Following section 4.2, both backbones generate chunks with $F = 3$ latent frames and $M = 880$ KV positions per frame, giving $FM = 2{,}640$ positions per native KV state. We use $M' = 440$, corresponding to a retention ratio $\rho = (M + M')/(FM) = 0.5$ and 1,320 retained positions per compressed historical chunk. We set the historical retrieval size to $K = 6$, preserve the initial state $\mathbf{B}_0$, and keep $R = 1$ recent native chunk for MG and $R = 2$ for ABot. The absorption parameters (α, τ_a) are $(0.25, 160)$ for MG and $(2.0, 140)$ for ABot. These settings preserve the native active attention length of each backbone; the full cache composition is given in Appendix A.1. ACE training details are provided in Appendix A.2.

5.2 Evaluation Setup

Evaluation Protocol. We evaluate three interaction horizons: *Short* (21 chunks), *Deep* (81 chunks), and *Complex* (113–117 chunks), together with a 241-chunk, *12-lap* trajectory for long-horizon analysis. MG uses 21 initial scenes with three trajectory variants, yielding 63 videos per horizon (189 total), while ABot uses nine `universal` scenes with three variants and three seeds, yielding 81 videos per horizon (243 total). Following WorldKV (Yi et al., 2026), revisits are paired by matching accumulated action-derived geometric states and scored only after the reference chunk leaves the native attention window ($t - h > 6$ for MG and $t - h > 7$ for ABot). We report PSNR, LPIPS, and SSIM over aligned revisit pairs. Detailed trajectories, pair counts, metric aggregation, and ABot scene selection are provided in Appendix B.

Baselines. We compare ConsistKV with Sliding, Full-KV, WorldKV (Yi et al., 2026), and LongLive-RAG (Hu et al., 2026). WorldKV and LongLive-RAG provide action-based and visual content-based retrieval baselines, respectively. Both retrieval baselines retain the same persistent initial chunk $\mathbf{B}_0$ as ConsistKV, while Sliding follows the backbone's native cache without a persistent anchor. For KV compression, we compare against WorldKV compression (Yi et al., 2026) and Forcing-KV compression (Ji et al., 2026), a recent KV compression method for autoregressive video diffusion models, under the same retained-position budget. Additional comparisons with generic KV compression and merging methods are provided in Appendix D. All retrieval-based methods use the same active KV budget, and all compression methods are matched to the same number of retained KV positions per historical chunk. Implementation details are provided in Appendix B.3.

5.3 Comparison with Existing Methods

Quantitative Comparison. As shown in Table 1, ConsistKV improves overall revisit consistency over fixed-budget retrieval baselines, with larger gains at longer interaction horizons. On MG, ConsistKV improves PSNR over WorldKV by $+0.33$, $+0.57$, and $+1.17$ dB on Short, Deep, and Complex, respectively, while also achieving the best average LPIPS and SSIM. Full-KV achieves a slightly higher average PSNR by retaining the full KV history, but reduces the generation speed from 26.9 to 16.1 FPS. On ABot, WorldKV performs slightly better on Short (14.61 vs. 14.36 dB), whereas ConsistKV improves PSNR on Deep and Complex to 14.62 and 14.97 dB, yielding the best three-horizon average (14.65 dB) among methods that complete all trajectories. Full-KV cannot execute the longer ABot trajectories within the available 178 GB GPU memory. An evaluation

Table 1: Quantitative comparison on Matrix-Game 2.0 and ABot-World. Short, Deep, and Complex report PSNR, while Avg. reports PSNR / LPIPS / SSIM over the three horizons. Best and second-best fidelity results are shown in **bold** and underline, respectively. ABot Full-KV† is evaluated only on Short, as Deep and Complex exceed 178 GB of GPU memory.

Method	Short PSNR↑	Deep PSNR↑	Complex PSNR↑	Avg. PSNR↑	Avg. LPIPS↓	Avg. SSIM↑	FPS↑
Matrix-Game 2.0							
Sliding (baseline)	12.28	10.51	10.47	11.09	0.616	0.240	33.2
Full-KV	**14.27**	14.48	**15.39**	**14.72**	0.523	0.391	16.1
LongLive-RAG (Hu et al., 2026)	12.49	11.78	11.01	11.76	0.575	0.284	23.8
WorldKV (Yi et al., 2026)	13.64	14.54	13.36	13.85	0.490	0.378	30.2
ConsistKV (Ours)	13.97	**15.11**	14.53	14.54	**0.479**	**0.409**	26.9
ABot-World							
Sliding (baseline)	13.86	13.04	13.88	13.59	0.656	0.497	16.8
Full-KV†	**14.97**	*OOM*	*OOM*	–	–	–	6.8
LongLive-RAG (Hu et al., 2026)	14.28	13.90	13.50	13.89	0.588	0.479	10.9
WorldKV (Yi et al., 2026)	14.61	14.34	13.83	14.26	0.568	0.487	11.3
ConsistKV (Ours)	14.36	**14.62**	**14.97**	**14.65**	**0.563**	**0.511**	10.8

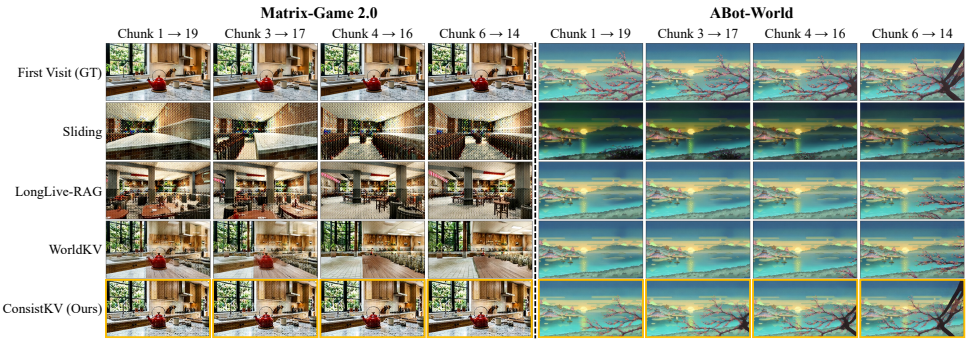


Figure 2: Qualitative comparison of revisit consistency on Matrix-Game 2.0 (left) and ABot-World (right). The first row shows the first-visit observation used as the reference (e.g., Chunk 1), and each subsequent row shows the frame generated by each method when revisiting the same accumulated geometric state (e.g., Chunk 19). Yellow borders denote ConsistKV.

on 20 additional initial images shows the same horizon-dependent trend, and full per-axis results are provided in Appendix C.

Qualitative Comparison. Figure 2 compares revisited frames with their corresponding first-visit observations. Sliding-window inference no longer has direct access to the first-visit KV state once it leaves the active context and shows substantial changes in scene layout and appearance. LongLive-RAG and WorldKV retain more of the previous scene structure, but mismatches remain in several examples. ConsistKV more closely matches the corresponding first-visit observations, preserving the kitchen layout and foreground objects on MG and the previously observed tree branches on ABot. These examples qualitatively support improved recall of distant historical states.

5.4 Analysis of ConsistKV Components

KV Compression. We first evaluate KV compression while fixing ACE retrieval and the number of retained positions to 1,320 per historical chunk. We compare against WorldKV (Yi et al., 2026) for world-model KV compression and Forcing-KV (Ji et al., 2026) for video-model KV compression under the same retained-position budget. As shown in the upper block of Table 2, WorldKV and Forcing-KV achieve average PSNRs of 13.93 and 13.75, respectively, under the same retained-position budget. Our Self-Gram KV Selection reaches 14.06 when the first latent frame is retained in full, compared with 13.92 without first-frame retention. Adding Gram-Kernel Value Absorption further improves the average PSNR to 14.54, while also achieving the best average LPIPS and SSIM. Relative to the same Self-Gram selection without absorption, the PSNR gains are +0.21, +0.21, and +1.01 dB on Short, Deep, and Complex, respectively. Additional comparisons with generic KV compression and merging methods are provided in Appendix D.

Table 2: Component analysis on Matrix-Game 2.0. The upper block compares KV compression with ACE retrieval fixed and 1,320 retained positions per historical chunk; the lower block compares retrieval strategies with Value-Absorbing KV Compression fixed. Short, Deep, and Complex report PSNR, while Avg. reports PSNR / LPIPS / SSIM over the three horizons. Best and second-best results are shown in **bold** and underline, respectively.

Method	Short PSNR↑	Deep PSNR↑	Complex PSNR↑	Avg. PSNR↑	Avg. LPIPS↓	Avg. SSIM↑
KV Compression (ACE retrieval fixed; 1,320 retained positions/chunk)						
WorldKV Compression (Yi et al., 2026)	13.59	14.56	13.64	13.93	0.486	0.384
Forcing-KV Compression (Ji et al., 2026)	13.41	14.32	13.51	13.75	0.492	0.374
Self-Gram KV Selection (w/o first-frame retention)	13.59	14.71	13.47	13.92	0.494	0.370
Self-Gram KV Selection	13.76	14.90	13.52	14.06	0.488	0.379
Value-Absorbing KV Compression (Ours)	**13.97**	**15.11**	**14.53**	**14.54**	**0.479**	**0.409**
Action–Content Retrieval (Value-Absorbing KV Compression fixed)						
LongLive-RAG (Content) (Hu et al., 2026)	13.37	13.39	12.85	13.20	0.547	0.343
WorldKV (Action) (Yi et al., 2026)	13.92	15.00	**14.56**	14.49	0.483	0.406
Action–Content Retrieval (ACE, Ours)	**13.97**	**15.11**	14.53	**14.54**	**0.479**	**0.409**

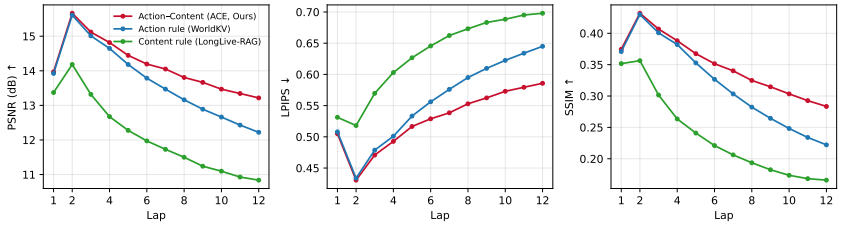


Figure 3: Long-horizon retrieval comparison on the 241-chunk, 12-lap trajectory. ACE remains close to action-only retrieval during early revisits, while its advantage becomes more visible as repeated visits accumulate.

Action–Content Retrieval. We compare ACE with content-based retrieval from LongLive-RAG (Hu et al., 2026) and action-based retrieval from WorldKV (Yi et al., 2026), while fixing Value-Absorbing KV Compression for all methods. As shown in the lower block of Table 2, ACE achieves the best three-horizon averages in PSNR, LPIPS, and SSIM. Its PSNR is slightly higher than action-only retrieval on Short and Deep (13.97 vs. 13.92 and 15.11 vs. 15.00), while the two are comparable on Complex (14.53 vs. 14.56). These results indicate that action-derived geometry already provides a strong localization cue on the main scripted trajectories.

The benefit of combining the two cues becomes clearer under repeated revisits. In figure 3, ACE and action-only retrieval remain close during early laps, while their performance gradually diverges as repeated revisits accumulate. This result suggests that visual content provides complementary information beyond action-derived geometry when retrieving historical states over long interactions. Although the gain of ACE over action-only retrieval is modest on the main benchmark, its advantage becomes more pronounced in the repeated long-horizon setting. Additional cross-path and retrieval analyses are provided in Appendix E.

6 Conclusion

We introduced *ConsistKV*, a lightweight long-term KV memory framework with action–content retrieval for mitigating revisit collapse in interactive world models. ConsistKV preserves compressed distant history outside the active cache and selectively recalls relevant states under a fixed attention budget. *Action–Content Retrieval* combines visual content with action-derived relative geometry to leverage complementary appearance and localization cues, enabling more reliable access to relevant historical states over long interactions. *Value-Absorbing KV Compression* reduces historical KV storage while preserving information from discarded positions through shared self-Gram-guided selection and value absorption. Experiments on Matrix-Game 2.0 and ABot-World show improved revisit consistency over sliding-window and fixed-budget memory baselines, with clearer gains at longer horizons and under repeated revisits. These results highlight the benefit of jointly designing retrieval and compression for persistent interactive generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, improving clarity and organization, and refining LaTeX presentation of author-written content. We did not use generative AI tools to generate experimental results, fabricate data, or make final scientific conclusions on behalf of the authors. All AI-assisted text and suggestions were manually reviewed, revised, and verified against the corresponding experiments, implementation, and cited literature by the authors. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies long-term memory and revisit consistency in generative interactive world models and does not involve human subjects or the collection of personal data. Our experiments use existing research models and publicly available or previously released evaluation images. We do not introduce a new dataset containing sensitive information. As with generative world models more broadly, improvements in long-horizon consistency could potentially be incorporated into downstream interactive generation systems; responsible deployment of such systems remains dependent on the application context. We are not aware of additional ethical concerns specific to the proposed memory mechanism.

REPRODUCIBILITY STATEMENT

We provide detailed information needed to reproduce our experiments throughout the main paper and appendix. The method section specifies the retrieval, KV compression, and memory-delivery procedures, while the appendix documents backbone-specific configurations, ACE architecture and training, mirror-retracing data construction, evaluation trajectories, revisit-pairing rules, baseline implementations, and hyperparameter settings. We additionally report detailed per-horizon results, compression and retrieval ablations, long-horizon retrieval analyses, and runtime settings. These details are intended to make both the proposed method and the evaluation protocol reproducible.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. In *International Conference on Learning Representations*, volume 2026, pp. 92075–92093, 2026.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, et al. Reworld: An interactive world model with long-horizon memory. *arXiv preprint arXiv:2608.23565*, 2026.

Harry Dong, Xinyu Yang, Zhenyu Zhang, Zhangyang Wang, Yuejie Chi, and Beidi Chen. Get more with less: Synthesizing recurrence with kv cache compression for efficient llm inference. *arXiv preprint arXiv:2402.09398*, 2024.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3d state. *arXiv preprint arXiv:2603.03482*, 2026.

Qiwen Gu, Bingjie Gao, Rui Chen, Geng Li, Jifan Li, Qishuai Wen, Li Niu, Jing Tang, Xiangxiang Chu, and Junqiao Zhao. R2m-bench: Evaluating revisit memory via relative consistency in interactive video world models. *arXiv preprint arXiv:2608.27328*, 2026.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

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.

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

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, XiTai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-kv: Hybrid kv cache compression for efficient autoregressive video diffusion models. *arXiv preprint arXiv:2605.09681*, 2026.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. Abot-world-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2(3):6, 2025.

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. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Zhan Peng, Jie Ma, Huiqiang Sun, Chong Gao, Zhijie Xue, Zhiyu Pan, Zhiguo Cao, Jun Liang, and Jing Li. Compression and retrieval: Implicit memory retrieval for video world models. *arXiv preprint arXiv:2606.23105*, 2026.

Dvir Samuel, Issar Tzachor, Matan Levy, Michael Green, Gal Chechik, and Rami Ben-Ari. Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. *arXiv preprint arXiv:2602.01801*, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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. *arXiv preprint arXiv:2512.14614*, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Zhongwei Wan, Xinjian Wu, Yu Zhang, Yi Xin, Chaofan Tao, Zhihong Zhu, Xin Wang, Siqi Luo, Jing Xiong, Longyue Wang, et al. D2o: Dynamic discriminative operations for efficient long-context inference of large language models. *arXiv preprint arXiv:2406.13035*, 2024.

Zheng Wang, Boxiao Jin, Zhongzhi Yu, and Minjia Zhang. Model tells you where to merge: Adaptive kv cache merging for llms on long-context tasks. *arXiv preprint arXiv:2407.08454*, 2024.

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

Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljoša Ošep. Addressable memory for video world models. *arXiv preprint arXiv:2608.07408*, 2026.

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

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M Patel, and Yiqun Mei. Wonder: Video world model done better. *arXiv preprint arXiv:2607.26037*, 2026.

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

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. Worldkv: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

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

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.

Jian Yuan, Ziwei He, Haoli Bai, Jingwen Leng, and Bo Jiang. Weightedkv: Attention scores weighted key-value cache merging for large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. Cam: Cache merging for memory-efficient llms inference. In *Forty-first international conference on machine learning*, 2024.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in neural information processing systems*, 36:34661–34710, 2023.

Lin Zhao, Yushu Wu, Yifan Gong, Yanzhi Wang, and Pu Zhao. Omnimem: Scalable and adaptive memory retrieval for long video generation. *arXiv preprint arXiv:2605.30519*, 2026a.

Wenqu Zhao, Xuemin Chi, Xin Zhang, Guoqing Ma, Baorun Li, Jianjie Fang, Peizhi Tang, Chen Gao, and Wei Wu. Densitykv: Density-guided kv cache compression for long video generation. *arXiv preprint arXiv:2608.27922*, 2026b.

A Implementation Details

A.1 Backbones and Memory Configuration

Following section 4.2, both backbones generate chunks with $F = 3$ latent frames and $M = 880$ KV positions per frame, so each native KV state contains $FM = 2{,}640$ positions. MG uses 30 layers, 12 attention heads, and per-head dimension $d = 128$. ConsistKV compresses only the video self-attention KV states. For Self-Gram KV Selection, the score g_j is computed independently in each head and summed across heads before selecting a shared position set $\mathcal{P}$; Gram-Kernel Value Absorption is then applied independently within each head.

We use $M' = 440$, so the retained state contains $M + M' = 1{,}320$ positions and the retention ratio is

$$\rho = \frac{M + M'}{FM} = 0.5. \tag{13}$$

Table 3 summarizes all backbone-specific hyperparameters used in the main experiments.

Table 3: ConsistKV configuration used in the main experiments. The total active positions include the chunk currently being generated and match the native attention length of each backbone.

	Matrix-Game 2.0	ABot-World
Native attention length	18 latent frames	21 latent frames
Native history L	5 chunks	6 chunks
KV positions per chunk FM	2,640	2,640
Retention M' / ρ	440 / 0.5	440 / 0.5
Retrieved history K	6 chunks	6 chunks
Recent native history R	1 chunk	2 chunks
Persistent anchor	$\mathbf{B}_0$	$\mathbf{B}_0$
Total active positions	15,840	18,480
Absorption (α, τ_a)	$(0.25, 160)$	$(2.0, 140)$
ACE trainable parameters	0.93M	1.04M

A.2 Action–Content Encoder

The content encoder E^{con} applies three strided convolutional layers with channel dimensions $128 \to 192 \to 256$, GroupNorm, and SiLU activations to a latent chunk $\mathbf{x}_h$. Global average pooling followed by a linear layer produces the 256-D descriptor $\mathbf{e}_h$. The fusion network $f(\mathbf{e}, \mathbf{s})$ first maps the 4-D relative geometric state $\mathbf{s}_{h \to t}$ to 32 dimensions with a two-layer MLP, concatenates it with $\mathbf{e}_h$, and maps the resulting 288-D feature to a 256-D embedding with another two-layer MLP followed by ℓ_2 normalization.

Mirror-Retracing training data. ACE is trained on generated mirror-retracing loops from eight `universal` images (`0000`–`0007`), disjoint from all quantitative evaluation images. For each training loop, the world model first generates a 14-chunk random outward walk and then generates a 14-chunk return walk by applying the inverse actions in reverse order. The return observations are therefore generated by the model rather than obtained by reversing the outward frames. This keeps the training queries within the same command-conditioned generation process used at inference, rather than treating reversed frames as observations. Table 4 summarizes the construction.

The 15 seeds are $7, 22, 55, 99, 202, 303, 404, 505, 606, 707, 808, 909, 1010, 1111$, and 1234. During ACE training, the candidate pool is restricted to historical chunks $h \in \{1, \ldots, t-7\}$ as implemented by the training builders, so supervision focuses on long-term retrieval. At deployment, ACE instead scores the historical storage $\mathcal{H}_t$ defined in section 4.1, excluding only the persistent anchor and the R recent native chunks.

Content-descriptor caching. Since $E^{\text{con}}(\mathbf{x}_h)$ is query-independent, $\mathbf{e}_h$ is computed once when chunk $\mathbf{x}_h$ enters historical storage and is cached together with its compressed KV state and geometric state. The original latent chunk therefore does not need to remain in memory.

Table 4: Mirror-Retracing training trajectories for ACE. Both backbones use the same recipe and the same 15 generation seeds.

	Matrix-Game 2.0	ABot-World
Training images	`universal 0000--0007`	same eight images
Action set	R, Lc, Cu, Cd, W, S, A, D, idle	W, A, S, D, I, J, K, L, idle
Outward walk	uniform random, 14 chunks	uniform random, 14 chunks
Return walk	reversed inverse actions, 14 generated chunks	reversed inverse actions, 14 generated chunks
Inverse mapping	W↔S, A↔D, R↔Lc, Cu↔Cd	W↔S, A↔D, J↔L, I↔K
Loop length	30 chunks	32 chunks
Number of loops	$8 \times 2 \times 15 = 240$	$8 \times 2 \times 15 = 240$
Positive h^+	most recent eligible historical chunk with matching $\mathbf{s}_h$	same
Clean-pair filter	latent cosine ≥ 0.45	latent cosine ≥ 0.45
Training	6,000 steps, batch 12, lr 3×10^{-4}, $\tau_c = 0.07$	same
Regularization	geometric-state jitter $\sigma = 0.3$, curriculum noise, content replacement $p = 0.1$	same

Table 5: Evaluation trajectories and scale. A chunk corresponds to one action held for three latent frames. "Pairs/video" is the number of scored out-of-window revisit pairs in each generated video, and "Videos" is the number of generated videos evaluated per method for each horizon.

				Pairs / video		Videos	
Horizon	Trajectory variants	Chunks	Construction	MG	ABot	MG	ABot
Short	Forward/backward, lateral strafe, yaw	21	initial idle + 8 outward + 2 idle + 8 inverse + 2 idle	5	8	63	81
Deep	Same three axis-aligned variants	81	corresponding 20-chunk round trip $\times 4$	65	68	63	81
Complex	Three mixed-action trajectories	113–117	mixed-action path executed twice	82–83	84–85	63	81
12-lap	Forward/backward, lateral strafe, yaw	241	corresponding 20-chunk round trip $\times 12$	225	–	63	–

Backbone-specific positional delivery. The positional adaptation $\mathcal{A}_t$ in equation 12 is implementation-dependent because MG and ABot store cached keys at different stages of RoPE (Su et al., 2024). MG (He et al., 2025) caches keys after RoPE has been applied, so relocation re-aligns only the temporal RoPE component of each retrieved key while leaving its spatial component and value unchanged. ABot (Jiang et al., 2026) caches raw keys and applies RoPE at attention time. For ABot, the original spatial coordinates of retained entries are stored as lightweight metadata and reused together with the reassigned temporal positions when applying RoPE. Neither adaptation introduces learned parameters.

B Evaluation Protocol

B.1 Initial Images and Trajectories

ACE training and quantitative evaluation use disjoint initial images. ACE is trained on eight `universal` images (`0000`–`0007`), as detailed in Appendix A.2. For the main benchmark, MG uses six `gta_drive`, six `temple_run`, and nine `universal` images (`0008`–`0016`) from the official Matrix-Game 2.0 repository (MIT license),[1] for 21 scenes in total. ABot uses the same nine `universal` evaluation images. MG uses generation seed 7, while ABot uses seeds 7, 99, and 123. We additionally evaluate ABot on 20 external initial images in Appendix C.2.

A chunk corresponds to three latent frames (12 decoded pixel frames) and one action input. The accumulated geometric state $\mathbf{s}_t$ is obtained only by integrating the issued commands; it does not use the motion actually realized by the generated video. We use three axis-aligned trajectory variants for Short: forward/backward, lateral strafe, and yaw. The forward/backward trajectory uses W/S, the lateral trajectory uses D/A, and the yaw trajectory uses R/Lc for MG and J/L for ABot. Deep and 12-lap repeat the corresponding Short round trip four and twelve times, respectively. Complex instead uses three mixed-action trajectories defined below. Table 5 summarizes both the trajectory construction and the number of evaluated videos.

The three Complex trajectories use mixed-action patterns $\mathcal{Q}_1$, $\mathcal{Q}_2$, and $\mathcal{Q}_3$. Importantly, the initial idle chunk is not part of the repeated pattern. Each trajectory has the form

$$[\,\cdot\,; \mathcal{Q}_i; \mathcal{Q}_i\,], \qquad i \in \{1, 2, 3\}, \tag{14}$$

[1] `https://github.com/SkyworkAI/Matrix-Game/tree/main/Matrix-Game-2/demo_images`

so the first pass creates the historical states and the second pass revisits the same accumulated geometric-state sequence after those states have left the native context. The three pattern lengths are 58, 56, and 58 chunks, giving total rollout lengths $1 + 2 \times 58 = 117$, $1 + 2 \times 56 = 113$, and $1 + 2 \times 58 = 117$, respectively. Their exact MG action patterns are:

$$\begin{aligned}
\mathcal{Q}_1 = {} & \mathrm{Lc}{\times}4\ \mathrm{S}{\times}3\ \mathrm{A}{\times}4\ \mathrm{D}{\times}2\ \mathrm{A}{\times}3\ \mathrm{S}{\times}2\ \mathrm{W}{\times}3\ \mathrm{Lc}{\times}2\ \mathrm{W}{\times}4\ \cdot{\times}2 \\
& \mathrm{S}{\times}4\ \mathrm{R}{\times}2\ \mathrm{S}{\times}3\ \mathrm{W}{\times}2\ \mathrm{D}{\times}3\ \mathrm{A}{\times}2\ \mathrm{D}{\times}4\ \mathrm{W}{\times}3\ \mathrm{R}{\times}4\ \cdot{\times}2, \\
\mathcal{Q}_2 = {} & \mathrm{R}{\times}2\ \mathrm{W}{\times}4\ \mathrm{Lc}{\times}3\ \mathrm{S}{\times}4\ \mathrm{R}{\times}2\ \mathrm{D}{\times}3\ \mathrm{S}{\times}2\ \mathrm{D}{\times}4\ \mathrm{S}{\times}2\ \cdot{\times}2 \\
& \mathrm{W}{\times}2\ \mathrm{A}{\times}4\ \mathrm{W}{\times}2\ \mathrm{A}{\times}3\ \mathrm{Lc}{\times}2\ \mathrm{W}{\times}4\ \mathrm{R}{\times}3\ \mathrm{S}{\times}4\ \mathrm{Lc}{\times}2\ \cdot{\times}2, \\
\mathcal{Q}_3 = {} & \mathrm{A}{\times}4\ \mathrm{W}{\times}4\ \mathrm{A}{\times}2\ \mathrm{R}{\times}3\ \mathrm{A}{\times}2\ \mathrm{Lc}{\times}4\ \mathrm{R}{\times}3\ \mathrm{S}{\times}3\ \mathrm{D}{\times}2\ \cdot{\times}2 \\
& \mathrm{A}{\times}2\ \mathrm{W}{\times}3\ \mathrm{Lc}{\times}3\ \mathrm{R}{\times}4\ \mathrm{D}{\times}2\ \mathrm{Lc}{\times}3\ \mathrm{D}{\times}2\ \mathrm{S}{\times}4\ \mathrm{D}{\times}4\ \cdot{\times}2.
\end{aligned}$$

For ABot, J/L are the yaw commands and I/K are the pitch commands; they are mapped to the same signed yaw and pitch updates used to construct the geometric state $\mathbf{s}_t$ for MG.

B.2 Revisit Pairing and Metrics

For evaluation, we denote the first-visit chunk by h to match the historical index used throughout section 4.1. A revisit pair (t, h) is created when $\mathbf{s}_h = \mathbf{s}_t$ exactly, using the earliest eligible historical chunk as the reference to measure consistency with the state established on the first visit. The initial image $\mathbf{x}_0$ is excluded. Only out-of-window pairs are scored, using $t - h > 6$ for MG and $t - h > 7$ for ABot.

For each scored pair, we compare three aligned decoded frames from chunks $\mathbf{x}_t$ and $\mathbf{x}_h$ (the central frame and offsets of ± 4 pixel frames) and average PSNR, LPIPS, and SSIM over the three frames. A video score is the mean over all scored revisit pairs, and each reported table entry is the mean over videos. The number of videos and scored pairs for each horizon is summarized in Table 5. Appendix E.2 additionally evaluates approximate pose-based revisits that do not require exact equality of the accumulated geometric state.

B.3 Baseline Implementations

Full-KV. Full-KV keeps the available KV history instead of evicting it. MG uses a 195-latent-frame cache for Short and 351 frames for Deep and Complex. On ABot, the Deep and Complex trajectories exceed the 178 GB available on one GPU, so only Short can be executed.

WorldKV. WorldKV (Yi et al., 2026) uses its action-based retrieval and KV compression under the same active attention budget as ConsistKV. In our implementation, its initial sink frames correspond to the first chunk $\mathbf{B}_0$, so WorldKV keeps the same persistent anchor as ConsistKV.

LongLive-RAG. LongLive-RAG (Hu et al., 2026) uses content-based retrieval and inserts three uncompressed historical chunks, matching the number of active KV positions used by six ConsistKV chunks at $\rho = 0.5$. The first chunk $\mathbf{B}_0$ likewise remains in the active cache together with the retrieved history. For ABot, whose latent dimensionality differs from the released retriever, we train the same retrieval architecture on ABot latents using training scenes disjoint from evaluation. Thus, all retrieval-based memory methods in our comparison use the same persistent anchor; native Sliding is the only baseline without it, and a same-budget Sliding+$\mathbf{B}_0$ control is reported in Table 10.

Forcing-KV. Forcing-KV (Ji et al., 2026) is used as a video-model KV-compression baseline. Its head-dependent hybrid compression applies structured compression to static heads and segment-wise similarity compression to dynamic heads. We adapt it to retain the same 1,320 KV positions per historical chunk as ConsistKV, while keeping ACE retrieval fixed.

WeightedKV. WeightedKV (Yuan et al., 2025) is re-implemented from the paper and adapted to the same 1,320-position budget with ACE retrieval fixed. It discards less-important keys while merging their associated values into retained neighboring positions using average-attention-score weights.

Table 6: Matrix-Game 2.0 per-axis and three-horizon average results. P/L/S denote PSNR / LPIPS / SSIM.

Method	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. PSNR↑	Avg. LPIPS↓	Avg. SSIM↑
ConsistKV (Ours)	13.97 / 0.475 / 0.401	15.11 / 0.464 / 0.415	14.53 / 0.498 / 0.413	14.54	**0.479**	**0.409**
Full-KV	14.27 / 0.502 / 0.389	14.48 / 0.549 / 0.360	15.39 / 0.518 / 0.424	**14.72**	0.523	0.391
WorldKV (Yi et al., 2026)	13.64 / 0.484 / 0.382	14.54 / 0.472 / 0.392	13.36 / 0.516 / 0.361	13.85	0.490	0.378
LongLive-RAG (Hu et al., 2026)	12.49 / 0.514 / 0.342	11.78 / 0.598 / 0.251	11.01 / 0.614 / 0.257	11.76	0.575	0.284
Sliding	12.28 / 0.524 / 0.322	10.51 / 0.671 / 0.179	10.47 / 0.653 / 0.218	11.09	0.616	0.240

Table 7: ABot-World per-axis and three-horizon average results. P/L/S denote PSNR / LPIPS / SSIM. Full-KV has no three-horizon average because only Short fits in memory.

Method	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. PSNR↑	Avg. LPIPS↓	Avg. SSIM↑
ConsistKV (Ours)	14.36 / 0.555 / 0.500	**14.62 / 0.561 / 0.514**	**14.97 / 0.574 / 0.520**	**14.65**	**0.563**	**0.511**
Full-KV	**14.97 / 0.518 / 0.507**	OOM	OOM	–	–	–
WorldKV (Yi et al., 2026)	14.61 / 0.539 / 0.494	14.34 / 0.564 / 0.490	13.83 / 0.601 / 0.477	14.26	0.568	0.487
LongLive-RAG (Hu et al., 2026)	14.28 / 0.562 / 0.486	13.90 / 0.590 / 0.483	13.50 / 0.613 / 0.467	13.89	0.588	0.479
Sliding	13.86 / 0.618 / 0.497	13.04 / 0.689 / 0.478	13.88 / 0.660 / 0.516	13.59	0.656	0.497

KVMerger. KVMerger (Wang et al., 2024) is likewise re-implemented from the paper and adapted to the same retained-position budget with ACE retrieval fixed. It identifies key-similar merging sets and jointly merges their key and value states using kernel-weighted aggregation.

B.4 ABot Evaluation Scenes

ABot does not interpret every control command as pure camera motion. In several GTA-driving and TempleRun scenes, translational inputs primarily move an on-screen agent, so applying the inverse command does not reliably return the generated viewpoint. We therefore use the nine `universal` scenes for the controlled main ABot revisit benchmark, where short out-and-back trajectories within the native context show substantially more reliable viewpoint return. This selection is used only to make the geometric revisit target well defined and does not involve any memory method. To verify that the main conclusion is not specific to these nine scenes, we additionally evaluate the same methods and protocol on 20 external initial images in Appendix C.2, comprising 60 generated videos per horizon and 180 videos across Short, Deep, and Complex for each method.

C Additional System Results

C.1 Full Per-Axis Results

The main table exposes per-axis PSNR to make the horizon-dependent trends explicit. Table 6 and Table 7 provide the corresponding PSNR / LPIPS / SSIM values for every axis together with the same three-horizon averages.

On MG, the PSNR gap between ConsistKV and WorldKV increases from 0.33 dB on Short to 0.57 dB on Deep and 1.17 dB on Complex. LongLive-RAG is lower across all three horizons, while Sliding shows a larger drop on Deep and Complex once the reference lies outside the local context. Overall, ConsistKV achieves the strongest three-horizon average among the fixed-budget retrieval methods.

ABot shows a complementary pattern. WorldKV is slightly stronger on Short, whereas ConsistKV is stronger on Deep and Complex and achieves the best three-horizon average among methods that complete all trajectories (14.65 vs. 14.26 dB PSNR for WorldKV). Full-KV achieves the highest Short PSNR but cannot be evaluated on the longer trajectories because of its memory growth.

C.2 Generalization to Additional Scenes

To further verify that the main ABot results are not specific to the nine `universal` evaluation scenes, we repeat the evaluation on 20 additional initial images outside the Matrix-Game 2.0 demo pool: 15 subject-free first-person images from WBench (MIT license), three Matrix-Game 3.0 demo images (Apache-2.0), and two YUME examples (Apache-2.0). These images are disjoint from ACE

Table 8: ABot results on 20 additional initial images. Each cell reports PSNR / LPIPS / SSIM. We evaluate $20 \times 3 = 60$ generated videos per horizon for each method, for 180 videos across Short, Deep, and Complex.

Method	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. PSNR↑	Avg. LPIPS↓	Avg. SSIM↑
ConsistKV (Ours)	13.94 / 0.522 / 0.464	**14.00 / 0.539 / 0.477**	**14.19 / 0.555** / 0.500	**14.04**	**0.539**	0.480
WorldKV (Yi et al., 2026)	**13.99 / 0.513** / 0.459	13.65 / 0.547 / 0.463	13.36 / 0.579 / 0.469	13.67	0.547	0.464
LongLive-RAG (Hu et al., 2026)	13.68 / 0.530 / 0.452	13.29 / 0.562 / 0.452	12.83 / 0.595 / 0.456	13.27	0.563	0.453
Sliding	12.78 / 0.588 / **0.466**	12.03 / 0.680 / 0.463	13.46 / 0.652 / **0.527**	12.76	0.640	**0.485**

Table 9: Detailed video KV compression comparison on MG with ACE retrieval fixed and 1,320 retained KV positions per historical chunk. Each cell reports PSNR / LPIPS / SSIM; the final column is the three-horizon average.

Compression Method	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. P/L/S
WorldKV Compression (Yi et al., 2026)	13.59 / 0.484 / 0.381	14.56 / 0.469 / 0.396	13.64 / 0.507 / 0.375	13.93 / 0.486 / 0.384
Forcing-KV Compression (Ji et al., 2026)	13.41 / 0.490 / 0.373	14.32 / 0.475 / 0.388	13.51 / 0.512 / 0.360	13.75 / 0.492 / 0.374
Self-Gram KV Selection (w/o first-frame retention)	13.59 / 0.488 / 0.372	14.71 / 0.471 / 0.388	13.47 / 0.522 / 0.350	13.92 / 0.494 / 0.370
Self-Gram KV Selection	13.76 / 0.483 / 0.382	14.90 / 0.464 / 0.398	13.52 / 0.517 / 0.357	14.06 / 0.488 / 0.379
Value-Absorbing KV Compression (Ours)	**13.97 / 0.475 / 0.401**	**15.11 / 0.464 / 0.415**	**14.53 / 0.498 / 0.413**	**14.54 / 0.479 / 0.409**

training and the main benchmark. We use the same trajectories, revisit-pairing rule, and ACE checkpoint as in the main evaluation. Full-KV is omitted because it cannot complete the longer horizons. Table 8 summarizes the results.

The additional scenes reproduce the horizon-dependent pattern of the main benchmark, providing an independent check beyond the nine `universal` scenes. ConsistKV and WorldKV are nearly tied on Short (13.94 vs. 13.99 dB), while ConsistKV improves PSNR by $+0.35$ dB on Deep and $+0.83$ dB on Complex, yielding a $+0.37$ dB three-horizon average gain. The same long-horizon trend is therefore observed across this broader set of initial appearances.

C.3 Additional Qualitative Results

Additional revisit examples. Figure 4 provides additional matched revisit examples on Matrix-Game 2.0, while Figure 5 shows the corresponding comparison on ABot-World. In both figures, the first-visit row is the earlier generated observation used as the revisit reference rather than external ground truth. Across these examples, ConsistKV shows closer agreement with previously established appearance and spatial structure when the same accumulated geometric state is revisited.

3rd-person long-horizon rollout quality. Although ConsistKV is designed primarily for revisit consistency, we additionally examine its behavior during general 3rd-person long-horizon rollouts. Figure 6 compares regularly sampled frames from ABot-World rollouts rather than matched revisit pairs. In the first example, several baselines gradually lose the foreground umbrella and drift toward substantially different scene content, whereas ConsistKV maintains the rainy urban setting and its salient foreground structure over the rollout. In the second example, the baselines exhibit stronger long-horizon degradation, including the character intersecting with scene geometry and severe visual corruption in later frames. ConsistKV instead maintains the 3rd-person character and surrounding old-town environment more consistently throughout the rollout. These qualitative results suggest that access to distant historical states may also support broader scene and object continuity beyond explicit revisits.

D KV Compression Analysis

D.1 Detailed KV Compression Comparison

Table 9 complements the compression block of Table 2 by adding LPIPS and SSIM for each horizon. All rows use the same ACE retrieval and retain 1,320 KV positions per historical chunk.

Forcing-KV, designed for autoregressive video diffusion, performs head-dependent KV pruning rather than transferring information from discarded values; under the matched retained-position budget, it reaches an average PSNR of 13.75. Retaining the first latent frame in full increases the

Table 10: Effect of the persistent anchor and the number of retrieved chunks on MG. The two Sliding controls use the native 18-frame (15,840-position) attention budget without retrieval; Sliding+$\mathbf{B}_0$ reserves the first chunk as a persistent anchor and uses the remaining slots for recent context. The K-sweep rows retain $\mathbf{B}_0$ and one recent native chunk, with $\rho = 1/2$ for each retrieved historical chunk. The $K = 0$ row is a reduced-context endpoint rather than a same-budget anchor control.

Configuration	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. P/L/S	Active positions
Sliding (no anchor)	12.28 / 0.524 / 0.322	10.51 / 0.671 / 0.179	10.47 / 0.653 / 0.218	11.09 / 0.616 / 0.240	15,840
Sliding + $\mathbf{B}_0$ anchor	13.15 / 0.515 / 0.351	13.05 / 0.582 / 0.286	12.01 / 0.603 / 0.282	12.74 / 0.566 / 0.306	15,840
$K = 0$ (no retrieval)	12.57 / 0.539 / 0.337	14.43 / 0.473 / 0.384	12.27 / 0.559 / 0.313	13.09 / 0.524 / 0.345	7,920
$K = 1$	12.96 / 0.507 / 0.356	13.73 / 0.497 / 0.365	11.78 / 0.583 / 0.296	12.83 / 0.529 / 0.339	9,240
$K = 2$	13.13 / 0.499 / 0.369	13.87 / 0.494 / 0.372	12.08 / 0.563 / 0.324	13.03 / 0.519 / 0.355	10,560
$K = 4$	13.69 / 0.481 / 0.389	14.47 / 0.482 / 0.395	13.49 / 0.525 / 0.382	13.88 / 0.496 / 0.389	13,200
$K = 6$ **(Ours)**	13.97 / 0.475 / 0.401	15.11 / 0.464 / 0.415	14.53 / 0.498 / 0.413	14.54 / 0.479 / 0.409	**15,840**
$K = 8$	13.84 / 0.494 / 0.379	14.94 / 0.485 / 0.389	14.81 / 0.488 / 0.435	14.53 / 0.489 / 0.401	18,480
$K = 10$	13.75 / 0.506 / 0.384	14.96 / 0.501 / 0.390	15.45 / 0.489 / 0.441	14.72 / 0.499 / 0.405	21,120

Table 11: Retention ratio ρ on MG with $K = 6$. Each P/L/S cell reports PSNR / LPIPS / SSIM.

ρ (positions/chunk)	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. P/L/S	Active positions
1/3 (880)	13.76 / 0.477 / 0.404	14.60 / 0.480 / 0.411	14.45 / 0.503 / 0.430	14.27 / 0.487 / 0.415	13,200
1/2 (1,320)**, Ours**	13.97 / 0.475 / 0.401	15.11 / 0.464 / 0.415	14.53 / 0.498 / 0.413	14.54 / 0.479 / 0.409	**15,840**
2/3 (1,760)	13.99 / 0.476 / 0.395	15.33 / 0.458 / 0.414	14.85 / 0.480 / 0.425	14.72 / 0.471 / 0.411	18,480
5/6 (2,200)	14.06 / 0.477 / 0.395	15.38 / 0.457 / 0.414	15.29 / 0.463 / 0.432	14.91 / 0.466 / 0.414	21,120
1 (2,640)	14.08 / 0.478 / 0.395	15.41 / 0.453 / 0.417	15.91 / 0.438 / 0.463	15.13 / 0.456 / 0.425	23,760

Self-Gram average PSNR from 13.92 to 14.06. Adding Gram-Kernel Value Absorption to the same first-frame-preserving selection further raises the average to 14.54, with PSNR gains of +0.21, +0.21, and +1.01 dB on Short, Deep, and Complex, respectively. The largest difference is observed on Complex.

D.2 Memory Allocation

We study how the active attention budget is divided between historical coverage and the number of KV positions retained for each retrieved location. Table 10 varies the number of retrieved chunks K at fixed compression $\rho = 1/2$, while Table 11 varies the retention ratio ρ at fixed $K = 6$. Each table reports PSNR / LPIPS / SSIM for every horizon together with the resulting active-cache size.

The persistent anchor provides a useful but incomplete memory signal. Under the same native active budget, adding $\mathbf{B}_0$ to Sliding raises the three-horizon average PSNR from 11.09 to 12.74, including a gain from 10.51 to 13.05 dB on Deep. The same control nevertheless remains below ConsistKV by 0.82, 2.06, and 2.52 dB on Short, Deep, and Complex, respectively. This comparison also clarifies that $K = 0$ is not a budget-matched anchor ablation: it uses only 7,920 active positions. Its relatively high Deep score (14.43 dB) is consistent with the Deep trajectory repeatedly returning to the initial region, where the persistent anchor is directly informative; the same $K = 0$ configuration remains 1.40 and 2.26 dB below ConsistKV on Short and Complex. The higher Deep score of $K = 0$ than Sliding+$\mathbf{B}_0$ should therefore not be interpreted as evidence that additional recent context is intrinsically harmful. In the reduced $K = 0$ cache, $\mathbf{B}_0$ is much more prominent relative to the remaining active context, which may increase reliance on the anchor on trajectories that repeatedly return near the start. We use Sliding+$\mathbf{B}_0$ as the budget-matched anchor control and $K = 0$ only as the zero-retrieval endpoint of the allocation sweep.

The K sweep is not strictly monotonic for very small retrieval sets. In particular, $K = 1$ and $K = 2$ can fall below $K = 0$ on Deep and Complex. This behavior is consistent with limited retrieval coverage when only one or two historical slots are available: on independent mirror-retracing queries, ACE reaches R@1 of 0.824 but R@6 of 1.000 without added geometric perturbation (Table 19). A missed top-ranked target can therefore occupy a large fraction of the retrieved context when K is small, whereas a larger retrieval set increases the chance that the relevant historical state is included. Performance recovers at $K = 4$ and is strongest among budget-matched settings at $K = 6$. Increasing to $K = 8$ or $K = 10$ exceeds the native MG attention budget; $K = 10$ further improves Complex and average PSNR but worsens average LPIPS and requires 21,120 active positions.

Table 12: Three allocations of the same 15,840-position active budget on MG with ACE retrieval fixed (63 videos per horizon). Each cell reports PSNR / LPIPS / SSIM; parentheses report the PSNR difference relative to ConsistKV.

Allocation	Short P/L/S	Deep P/L/S	Complex P/L/S
$\rho = 1/2$, $K = 6$ **(Ours)**	**13.97 / 0.475 / 0.401**	**15.11 / 0.464 / 0.415**	**14.53 / 0.498 / 0.413**
$\rho = 3/4$, $K = 4$	13.68 (-0.29) / 0.478 / 0.387	14.61 (-0.50) / 0.475 / 0.394	13.36 (-1.17) / 0.522 / 0.373
Uncompressed, $K = 3$	13.39 (-0.58) / 0.487 / 0.374	13.72 (-1.39) / 0.495 / 0.362	12.83 (-1.70) / 0.520 / 0.340

Table 13: Gram-Kernel Value Absorption sweep on MG (three horizons, 63 videos per horizon, one generation seed). Each cell reports PSNR / LPIPS / SSIM; the last column is the three-horizon mean.

Setting	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. P/L/S
$\alpha = 0$ (no absorption)	13.76 / 0.483 / 0.382	14.90 / 0.464 / 0.398	13.52 / 0.517 / 0.357	14.06 / 0.488 / 0.379
$\alpha = 0.25, \tau_a = 80$	13.88 / 0.474 / 0.400	15.03 / 0.462 / 0.418	14.14 / 0.501 / 0.410	14.35 / 0.479 / 0.409
$\alpha = 0.25, \tau_a = 160$ **(Ours)**	13.97 / 0.475 / 0.401	15.11 / 0.464 / 0.415	14.53 / 0.498 / 0.413	14.54 / 0.479 / 0.409
$\alpha = 0.25, \tau_a = 240$	13.96 / 0.476 / 0.400	15.09 / 0.465 / 0.413	14.54 / 0.501 / 0.411	14.53 / 0.481 / 0.408
$\alpha = 0.5, \tau_a = 160$	13.74 / 0.481 / 0.415	14.39 / 0.499 / 0.425	15.23 / 0.484 / 0.486	14.45 / 0.488 / 0.442
$\alpha = 1.0, \tau_a = 160$	12.78 / 0.530 / 0.440	12.25 / 0.594 / 0.443	15.00 / 0.465 / 0.588	13.34 / 0.530 / 0.490

Keeping $K = 6$ fixed, increasing ρ improves per-location fidelity because each retrieved chunk retains more of its original KV state, but the active cache grows proportionally. The uncompressed setting reaches the highest fidelity in this sweep but requires 23,760 active positions, $1.5\times$ the native MG budget. We use $\rho = 1/2$ to keep six retrieved locations within the original attention length, with Gram-Kernel Value Absorption applied to the resulting compressed states.

To isolate this coverage–fidelity trade-off under an exactly matched active budget, Table 12 compares three allocations that all use 15,840 active positions.

Relative to three uncompressed locations, the PSNR gain widens from $+0.58$ dB on Short to $+1.39$ dB on Deep and $+1.70$ dB on Complex, while the intermediate allocation with $\rho = 3/4$ and $K = 4$ yields intermediate PSNR across all three horizons.

D.3 Absorption Strength

We analyze the sensitivity of Gram-Kernel Value Absorption within Value-Absorbing KV Compression. Both sweeps use generation seed 7 and keep the retrieval, retention, and active-budget settings fixed to the main configuration; only the absorption strength α and temperature τ_a are varied. In both tables, $\alpha = 0$ corresponds to first-frame-preserving Self-Gram KV Selection without value absorption. Table 13 and Table 14 report results for all three horizons together with their three-horizon mean.

Relative to selection without absorption ($\alpha = 0$), our MG setting improves the three-horizon mean PSNR from 14.06 to 14.54 dB, while reducing LPIPS from 0.488 to 0.479 and increasing SSIM from 0.379 to 0.409. At fixed $\alpha = 0.25$, increasing τ_a from 80 to 160 improves PSNR by $+0.09$, $+0.08$, and $+0.39$ dB on Short, Deep, and Complex, respectively, whereas increasing it further to 240 leaves performance nearly unchanged. More aggressive absorption shifts the trade-off across horizons: $\alpha = 0.5$ improves Complex but lowers Short and Deep, and $\alpha = 1.0$ further degrades the shorter horizons. We therefore use $(\alpha, \tau_a) = (0.25, 160)$ as a stable setting across interaction horizons.

Relative to selection without absorption, our ABot setting improves the three-horizon mean PSNR from 14.24 to 14.87 dB, reduces LPIPS from 0.571 to 0.562, and increases SSIM from 0.485 to 0.517. At fixed $\alpha = 2.0$, increasing τ_a from 70 to 140 leaves Deep nearly unchanged (14.83 vs. 14.84 dB) but improves Short from 14.44 to 14.57 dB and Complex from 14.68 to 15.20 dB, yielding a $+0.22$ dB gain in the three-horizon mean. At $\tau_a = 140$, increasing α from 0.5 to 2.0 progressively improves the mean PSNR from 14.43 to 14.87 dB, driven mainly by stronger Deep and Complex performance. Increasing α further to 4.0 raises Complex to 15.54 dB and SSIM to 0.531, but reduces Short and Deep and worsens average LPIPS to 0.585. We therefore use $(\alpha, \tau_a) = (2.0, 140)$ as a balanced setting across the three horizons, achieving the highest mean PSNR and lowest mean LPIPS among the tested settings while retaining strong SSIM.

Table 14: Gram-Kernel Value Absorption sweep on ABot (three horizons, 27 videos per horizon, one generation seed). Each cell reports PSNR / LPIPS / SSIM; the last column is the three-horizon mean.

Setting	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. P/L/S
$\alpha = 0$ (no absorption)	14.63 / 0.540 / 0.488	14.50 / 0.564 / 0.494	13.59 / 0.610 / 0.472	14.24 / 0.571 / 0.485
$\alpha = 2.0, \tau_a = 70$	14.44 / 0.553 / 0.499	14.83 / 0.557 / 0.518	14.68 / 0.579 / 0.500	14.65 / 0.563 / 0.506
$\alpha = 0.5, \tau_a = 140$	14.49 / 0.550 / 0.489	14.50 / 0.563 / 0.501	14.31 / 0.594 / 0.499	14.43 / 0.569 / 0.496
$\alpha = 1.0, \tau_a = 140$	14.59 / 0.554 / 0.499	14.80 / 0.557 / 0.511	14.70 / 0.586 / 0.501	14.70 / 0.566 / 0.504
$\alpha = 2.0, \tau_a = 140$ **(Ours)**	14.57 / 0.555 / 0.507	14.84 / 0.558 / 0.521	15.20 / 0.572 / 0.522	14.87 / 0.562 / 0.517
$\alpha = 4.0, \tau_a = 140$	14.49 / 0.579 / 0.514	14.50 / 0.595 / 0.523	15.54 / 0.582 / 0.557	14.84 / 0.585 / 0.531

Table 15: Additional KV compression and merging baselines with ACE retrieval and 1,320 retained KV positions per historical chunk on MG (63 videos per horizon). Each cell reports PSNR / LPIPS / SSIM. † denotes methods re-implemented from the paper descriptions because public code was unavailable; implementation details are provided in Appendix B.

Compression Method	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. PSNR↑	Avg. LPIPS↓	Avg. SSIM↑
H2O (Zhang et al., 2023)	13.74 / 0.481 / 0.384	14.84 / 0.463 / 0.400	13.89 / 0.508 / 0.365	14.16	0.484	0.383
SnapKV (Li et al., 2024)	13.71 / 0.482 / 0.381	14.87 / 0.460 / 0.400	13.95 / 0.502 / 0.368	14.18	0.481	0.383
PyramidKV (Cai et al., 2024)	13.90 / 0.477 / 0.387	15.13 / 0.451 / 0.413	14.44 / 0.494 / 0.386	<u>14.49</u>	**0.474**	0.395
WeightedKV (Yuan et al., 2025)†	13.77 / 0.482 / 0.378	14.93 / 0.462 / 0.398	14.02 / 0.506 / 0.378	14.24	0.483	0.385
KVMerger (Wang et al., 2024)†	13.44 / 0.508 / 0.414	13.70 / 0.521 / 0.428	14.42 / 0.448 / 0.499	13.85	0.492	**0.447**
Value-Absorbing KV Compression (Ours)	13.97 / 0.475 / 0.401	15.11 / 0.464 / 0.415	14.53 / 0.498 / 0.413	**14.54**	<u>0.479</u>	<u>0.409</u>

D.4 Additional Compression Baselines

We additionally compare H2O, SnapKV, PyramidKV, WeightedKV, and KVMerger under the same ACE retrieval while retaining 1,320 KV positions per historical chunk in Table 15. H2O, SnapKV, and PyramidKV use an 880-position observation window; for PyramidKV, the official text-model default of 64 would allocate 2,514 positions to the first layer and exceed the native MG attention budget. WeightedKV and KVMerger are re-implemented from their paper descriptions and adapted to the same video-KV setting and retained-position budget.

PyramidKV achieves the best average LPIPS, while our Value-Absorbing KV Compression achieves the highest average PSNR. WeightedKV and KVMerger provide particularly relevant comparisons because both preserve information from discarded KV states rather than relying on pruning alone. WeightedKV uses attention-derived importance to determine retention and redistributes discarded-value information to retained neighboring states, whereas KVMerger forms key-similar groups and jointly merges their key and value states. In contrast, our method reuses a single self-Gram structure for both representative-position selection and soft value absorption, directly coupling the two compression stages. Under the same retained-position budget, our method achieves an average PSNR of 14.54 dB, compared with 14.24 for WeightedKV and 13.85 for KVMerger. PyramidKV achieves the best average LPIPS (0.474), while KVMerger achieves the best average SSIM (0.447). These results support the effectiveness of our compression method.

E Action–Content Retrieval Analysis

E.1 Detailed Retrieval Comparison

Table 16 complements the retrieval block of Table 2 by reporting PSNR / LPIPS / SSIM for every trajectory and the three-horizon average.

Action Retrieval is 0.03 dB higher on Complex PSNR alone (14.56 vs. 14.53), whereas ACE is higher on Short and Deep and yields the stronger three-horizon average (14.54 vs. 14.49), together with better average LPIPS and SSIM. Action-derived geometry already provides a strong localization cue on the scripted trajectories, while visual content supplies complementary appearance information and yields a modest aggregate improvement across the three horizons. The following cross-path, controlled-noise, and repeated-lap analyses further examine the two cues under different retrieval conditions.

Table 16: Detailed retrieval comparison on MG with Value-Absorbing KV Compression fixed. Each cell reports PSNR / LPIPS / SSIM.

Retrieval Method	Short P/L/S	Deep P/L/S	Complex P/L/S	Avg. P/L/S
Content Retrieval (LongLive-RAG) (Hu et al., 2026)	13.37 / 0.496 / 0.383	13.39 / 0.555 / 0.320	12.85 / 0.589 / 0.325	13.20 / 0.547 / 0.343
Action Retrieval (WorldKV) (Yi et al., 2026)	13.92 / 0.477 / 0.398	15.00 / 0.470 / 0.410	**14.56** / 0.501 / 0.410	14.49 / 0.483 / 0.406
Action–Content Retrieval (ACE, Ours)	**13.97 / 0.475 / 0.401**	**15.11 / 0.464 / 0.415**	14.53 / **0.498 / 0.413**	**14.54 / 0.479 / 0.409**

Table 17: Cross-path trajectories used on MG. $\cdot$ denotes an idle chunk and $\mathrm{K} \times n$ denotes holding command K for n consecutive chunks. All trajectories begin with one idle chunk.

Trajectory structure	Chunks	Command sequence
Full rotation	18	$\cdot$ R×14 $\cdot$ ×3
Forward–turn revisit	26	$\cdot$ W×4 R×7 W×4 R×7 W×2 $\cdot$
Random walk I	90	$\cdot$ A×6 S×9 Cd×5 Cu×3 S×5 W×5 Cu×6 W×5 R×3 S×4 W×13 D×5 W×6 D×5 Cd×6 D×3
Random walk II	90	$\cdot$ W×2 S×4 A×4 R×6 D×6 W×6 A×5 Cu×6 Lc×6 Cd×6 R×2 W×4 Cd×4 Cu×5 A×9 D×2 A×7 Lc×5
Random walk III	90	$\cdot$ D×6 A×4 Cd×6 S×6 W×5 R×6 D×3 Cd×11 A×9 Cu×2 S×3 W×8 Cd×6 Cu×11 Cd×3
Turn-detour return	33	$\cdot$ W×8 Lc×3 W×5 S×5 R×3 S×8
Multi-turn return	41	$\cdot$ W×6 R×4 W×6 R×4 Lc×4 S×6 Lc×4 S×6
Strafe-detour return	37	$\cdot$ W×8 D×6 W×4 S×4 A×6 S×8

E.2 Cross-Path Revisit Evaluation

The main benchmark pairs revisits by exact equality of the accumulated geometric state, which primarily evaluates retraced trajectories. We additionally consider eight MG trajectories that revisit earlier regions without simply repeating the same action sequence, including full rotations, forward–turn revisits, random walks, and outward-and-return paths with turning or strafing detours. The exact command sequences are summarized in Table 17.

A revisit is defined using the accumulated geometric state from section 4.1: the planar distance between two chunks must be below 0.6 and their wrapped yaw difference below 0.25 rad. If multiple historical chunks satisfy these conditions, we use the earliest eligible historical chunk as the reference. As in the main evaluation, only out-of-window pairs with $t - h > 6$ are scored. Across 21 scenes and eight trajectories, this yields 168 rollouts per method. Value-Absorbing KV Compression is fixed for all retrieval methods, while Sliding is included as a no-memory reference.

As shown in Table 18, ACE achieves the best aggregate PSNR, LPIPS, and SSIM across the eight trajectories. Compared with action-only and content-only retrieval, ACE improves PSNR by 0.48 and 0.49 dB, respectively. Sliding remains competitive in aggregate PSNR, while its LPIPS and SSIM remain lower than those of the memory-based methods. These results provide additional evidence that visual content and action-derived geometry contribute complementary information across a broader set of revisit paths.

E.3 Independent Retrieval Accuracy

We directly evaluate retrieval on 48 independent mirror-retracing loops generated with three different seeds on MG. For each return query, the candidate set contains out-of-window chunks $1, \ldots, t - 7$, and the positive is the most recent chunk with the same accumulated geometric state; noisy pairs with latent cosine below 0.45 are excluded. We compare WorldKV-style action retrieval based on relative geometry, LongLive-RAG-style content retrieval using latent cosine similarity, and ACE. As a controlled stress test, we add Gaussian perturbations with standard deviation σ to the query geometric state to examine whether visual content remains useful when the geometric cue becomes imperfect; this is not intended to model a specific source of trajectory drift. Table 19 reports Recall@K over 290 queries; $K = 6$ matches the retrieval size used at inference.

Action retrieval is already strong without added perturbation, reaching perfect R@6 at $\sigma = 0$. As the geometric cue is perturbed, ACE degrades more slowly and achieves the highest R@1 and R@6 across all nonzero noise levels. At $\sigma = 2$, ACE reaches R@6 of 0.740, compared with 0.668 for action retrieval and 0.641 for content retrieval. This controlled stress test supports the complementary role of visual content when the geometric cue is perturbed. Since the positive labels are defined by accumulated commands rather than measured physical pose, the experiment should be interpreted as a controlled robustness analysis rather than absolute localization accuracy.

Table 18: Cross-path revisit evaluation on MG (21 scenes $\times$ 8 trajectories $=$ 168 rollouts). Results are aggregated over all trajectories.

Retrieval	PSNR↑	LPIPS↓	SSIM↑
Action–Content Retrieval (ACE, Ours)	**13.98**	**0.543**	**0.362**
Action Retrieval (WorldKV) (Yi et al., 2026)	13.50	0.559	0.348
Content Retrieval (LongLive-RAG) (Hu et al., 2026)	13.49	0.566	0.336
Sliding (no memory)	13.72	0.585	0.293

Table 19: Independent retrieval accuracy on MG under controlled geometric-state perturbations. R@K denotes the fraction of queries for which the target historical chunk appears within the top K candidates. σ denotes the standard deviation of Gaussian noise added to the query geometric state.

	Action (WorldKV)		Content (LongLive-RAG)		ACE (Ours)	
Geometric-state noise σ	R@1	R@6	R@1	R@6	R@1	R@6
0.0	0.714	**1.000**	0.134	0.641	**0.824**	**1.000**
0.5	0.456	0.969	0.134	0.641	**0.554**	**0.971**
1.0	0.249	0.833	0.134	0.641	**0.338**	**0.864**
2.0	0.143	0.668	0.134	0.641	**0.207**	**0.740**
3.0	0.111	0.597	0.134	0.641	**0.180**	**0.682**

E.4 Long-Horizon Retrieval

The 12-lap trajectory repeats the same route for 241 chunks while Value-Absorbing KV Compression is fixed. Table 20 reports all three metrics at selected revisit laps and their mean over the 12 laps.

WorldKV, which denotes Action Retrieval (Yi et al., 2026), is nearly indistinguishable from ACE at the first revisit, but the gap widens as the route is revisited repeatedly. By Lap 12, ACE is higher by 1.00 dB PSNR and 0.061 SSIM, with 0.059 lower LPIPS. LongLive-RAG, which denotes Content-only Retrieval (Hu et al., 2026), shows a larger decline over the repeated laps. This trend suggests that visual content provides complementary information beyond action-derived geometry as repeated revisits accumulate.

Qualitative long-horizon revisits. Figure 7 visualizes repeated revisits over increasing temporal gaps on Matrix-Game 2.0, while Figure 8 shows the corresponding long-horizon behavior on ABot-World. In both cases, the same accumulated geometric state is revisited after progressively longer rollouts, and the qualitative differences between ConsistKV and the fixed-budget baselines become clearer at later revisits.

F Runtime and Memory Analysis

Measurement protocol. FPS in Table 1 measures generation speed and is evaluated separately from revisit fidelity. We use a dedicated 31-chunk runtime-only trajectory consisting of one initial idle chunk followed by three repetitions of 5 right-yaw and 5 left-yaw chunks; this trajectory is not used for Short, Deep, Complex, or 12-lap evaluation. Runtime is measured on an otherwise idle machine with a single NVIDIA B200 GPU using sequential execution and bf16. VAE decoding and metric computation are excluded. We report the wall-clock FPS of the final generated chunk, averaged over two scenes and two repeated runs. The MG methods are measured in the same session, and differences of roughly 1–1.5 FPS should be treated as measurement variation.

As shown in Table 21, ConsistKV preserves the same fixed active attention budget as WorldKV and LongLive-RAG, while Full-KV expands the attended history as the rollout grows. ConsistKV retains 1,320 KV positions per historical chunk, matching WorldKV and requiring half the external KV storage of the uncompressed LongLive-RAG configuration. At a 241-chunk rollout, this corresponds to an analytical external bank of 57.2 GB for ConsistKV and WorldKV, compared with 114.4 GB for LongLive-RAG. In our long-rollout implementation, external banks beyond 200 historical chunks

Table 20: Long-horizon retrieval on MG. We report PSNR, LPIPS, and SSIM at selected revisit laps and their mean over all 12 laps. Best and second-best results are shown in **bold** and underline, respectively.

Lap	Metric	**ConsistKV (Ours)**	WorldKV (Yi et al., 2026)	LongLive-RAG (Hu et al., 2026)
1	PSNR↑	**13.97**	13.92	13.37
	LPIPS↓	**0.505**	0.508	0.531
	SSIM↑	**0.374**	0.371	0.352
4	PSNR↑	**14.82**	14.65	12.67
	LPIPS↓	**0.493**	0.501	0.603
	SSIM↑	**0.388**	0.382	0.263
8	PSNR↑	**13.81**	13.16	11.50
	LPIPS↓	**0.553**	0.595	0.673
	SSIM↑	**0.325**	0.282	0.194
12	PSNR↑	**13.22**	12.22	10.84
	LPIPS↓	**0.586**	0.645	0.698
	SSIM↑	**0.283**	0.222	0.166
Mean	PSNR↑	**14.15**	13.67	12.09
	LPIPS↓	**0.528**	0.558	0.633
	SSIM↑	**0.348**	0.318	0.235

Table 21: Runtime and memory characteristics on MG under the main experimental configuration. Active KV positions include the chunk currently being generated. Historical KV/chunk reports the number of positions retained for each historical chunk relative to its native 2,640 positions. The 241-chunk bank size is analytical and excludes model weights.

Method	Active KV positions	Historical KV / chunk	External bank at 241 chunks	FPS↑
Sliding	15,840	–	–	33.2
Full-KV	grows with history	full history in active cache	–	16.1
LongLive-RAG (Hu et al., 2026)	15,840	2,640 (1.0×)	114.4 GB	23.8
WorldKV (Yi et al., 2026)	15,840	1,320 (0.5×)	57.2 GB	30.2
ConsistKV (Ours)	15,840	1,320 (0.5×)	57.2 GB	26.9

are kept in host memory; the 241-chunk values in the table therefore describe the stored KV bank rather than peak GPU allocation.

Under the clean-machine protocol above, ConsistKV runs at 26.9 FPS on MG, compared with 30.2 FPS for WorldKV, 23.8 FPS for LongLive-RAG, and 16.1 FPS for Full-KV. The remaining overhead relative to WorldKV includes the additional Value-Absorbing KV Compression performed when a chunk is committed to historical memory. ACE operates on cached 256-D content descriptors, so historical latent chunks are not re-encoded at every retrieval step.

G LIMITATIONS

ConsistKV bounds the active attention cost by retrieving only a fixed number of historical KV states, but it does not bound the total size of the external memory. Although each historical chunk is compressed by a factor of ρ, the compressed history still grows as $\mathcal{O}(\rho T)$ with rollout length. Under the main MG configuration, the external bank reaches 57.2 GB at 241 chunks, matching WorldKV and remaining half the size of the uncompressed LongLive-RAG bank (Appendix F). In our long-rollout implementation, banks beyond 200 historical chunks are kept in host memory. Thus, the fixed active attention budget should not be interpreted as bounded total memory usage, and substantially longer or persistent deployments may benefit from memory consolidation, eviction, hierarchical storage, or more memory-efficient residency policies.

Our formulation also primarily targets revisits to previously established scene states. Mirror-Retracing therefore assumes that an earlier visual state remains a meaningful reference when the

same location is revisited. This assumption can become less appropriate in strongly dynamic environments, where objects, agents, or other scene content may legitimately change over time. In such cases, long-term memory should retain persistent information while also accounting for temporally evolving state, rather than simply encouraging agreement with an earlier observation.



Figure 4: Additional qualitative revisit examples on Matrix-Game 2.0. Each column shows a first-visit/revisit chunk pair. The first row is the earlier generated observation used as the reference, and the remaining rows show the corresponding revisit generated by each method. Red boxes highlight local content changes and yellow borders denote ConsistKV.



Figure 5: Additional qualitative revisit examples on ABot-World. Each column compares a revisited state with its earlier generated reference. ConsistKV shows closer agreement with scene-specific objects and layout in the highlighted revisit pairs; red boxes indicate representative differences and yellow borders denote ConsistKV.



Figure 6: Qualitative comparison during a general 3rd-person long-horizon rollout on ABot-World. Frames are sampled from the same 81-chunk rollout at Chunks 1, 17, 33, 49, 65, and 81 rather than from matched revisit pairs. The examples suggest a potential benefit of historical memory for maintaining scene continuity beyond explicit revisits.

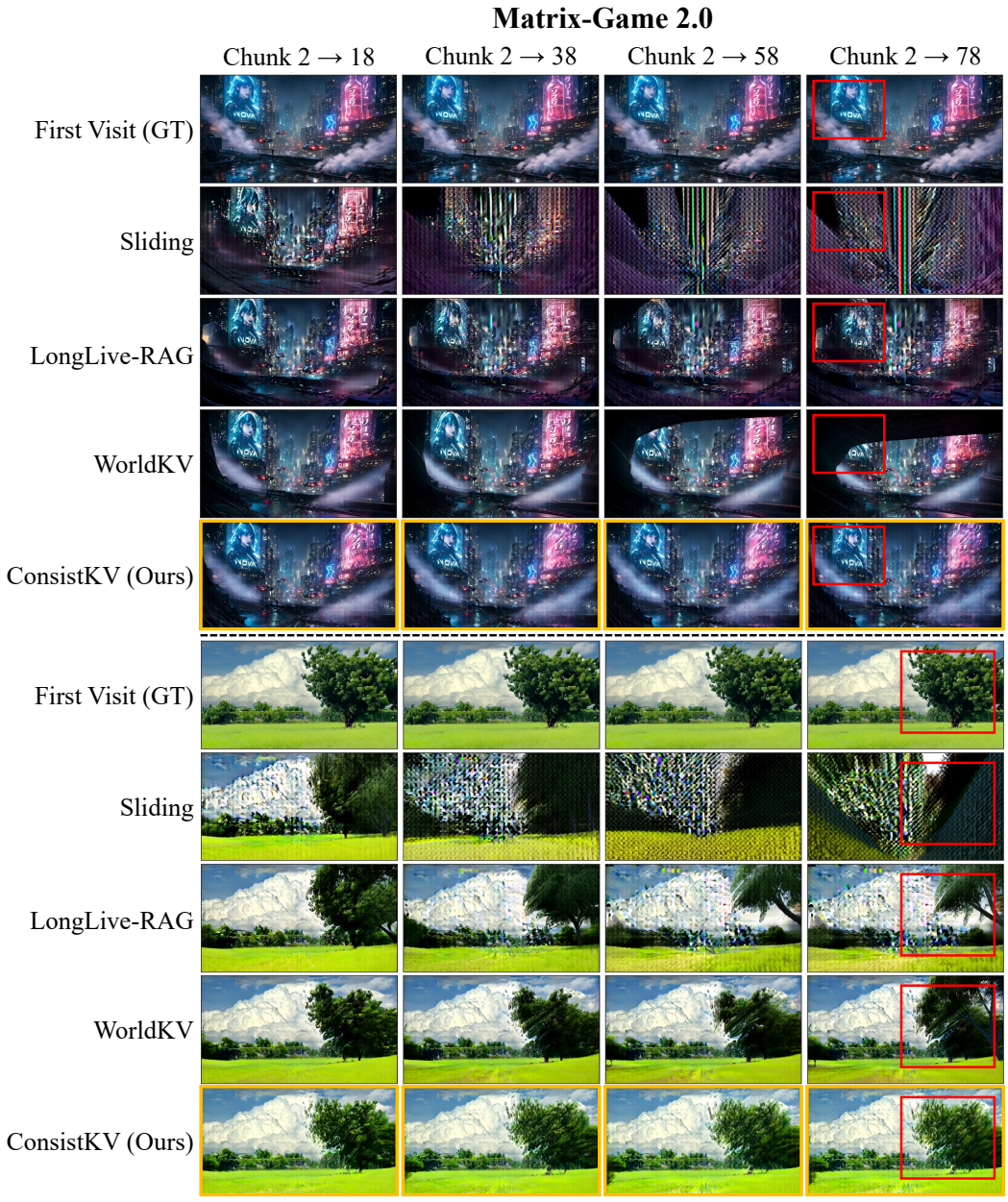


Figure 7: Repeated long-horizon revisit comparison on Matrix-Game 2.0. The accumulated geometric state of the first-visit reference at Chunk 2 is revisited after progressively longer rollouts (Chunks 18, 38, 58, and 78). ConsistKV shows closer agreement with the earlier scene content at later revisits; red boxes highlight representative changes and yellow borders denote ConsistKV.



Figure 8: Repeated long-horizon revisit comparison on ABot-World. The accumulated geometric state of the first-visit reference at Chunk 2 is revisited at increasing temporal gaps. In these examples, ConsistKV remains closer to the earlier visual state at later revisits, while the fixed-budget baselines show larger differences from the reference.