PERSISTENT MEMORY BRIDGES OCCLUSIONS IN VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory-based video object segmentation models such as SAM2 condition each frame on a fixed-size window of recent memory frames. We show that this fixed memory horizon is a substantial, previously unaddressed cause of a costly failure: once an object is occluded for longer than the window, every pre-occlusion memory frame has been evicted and the model falls back on a stale initialization it rarely recovers from. We introduce Persistent Memory Reinjection (PMR), a training-free mechanism that keeps a small buffer of pre-occlusion memory alive past the window and reinjects it into the model's own cross-attention when the recency window is incomplete, an exact and free occlusion signal whenever tracking is interrupted; for unconstrained video, where memory is written on every frame, we evaluate always-on and confidence-triggered variants of the same mechanism. On a controlled occlusion benchmark over 1500 MOSEv2 sequences, PMR improves post-occlusion region accuracy by up to 17 $\mathcal{J}$ points (up to 6 on DAVIS-2017), bootstrap-significant at every occlusion length beyond the window, while leaving continuous-tracking accuracy exactly unchanged. Using only the four freshest pre-occlusion frames, PMR matches an infinite-horizon oracle that attends to all past memory, showing that of the re-acquisition accuracy recoverable without new supervision, the bounded horizon rather than memory content or capacity is the main limiting factor. PMR composes with fine-tuning, which improves continuous tracking but not re-acquisition, and extends to naturally occurring failures of several types.

1 INTRODUCTION

Video object segmentation (VOS) asks a model to track and segment a target object across every frame of a video given only its mask in the first frame. The dominant modern approach, exemplified by SAM2 (Ravi et al., 2024), augments an image segmenter with a memory: as it tracks, it encodes each frame's prediction into a memory representation and conditions future frames on a set of recent memories through cross-attention. This memory is what lets the model maintain a consistent notion of which object to segment over time. The central open problem in this paradigm is long-horizon robustness: after an object is occluded, leaves the frame, or is otherwise lost for an extended period, memory-based trackers frequently fail to re-acquire it, and the failure is often permanent. Prior work has attributed such failures to drift in the memory representation and has sought better memory features or learned re-identification modules, but these interventions yield small and inconsistent gains on the cases that matter (Cheng & Schwing, 2022; Yang et al., 2021).

We instead attribute a large fraction of long-occlusion failures not by what the memory contains but by how far back it reaches. SAM2 selects the memories it conditions on by recency: it always attends to the most recent m tracked frames (plus the first frame). When an occlusion lasts longer than m frames, none of the frames inside the occlusion were tracked, so the recency window contains no usable memory and the model is left conditioning on the first-frame template alone, whose appearance has by then drifted from the reappearing object. Our contribution is Persistent Memory Reinjection (PMR), a simple and training-free mechanism built on this observation. PMR keeps a small first-in-first-out buffer of recent memory frames that, unlike the attention window, is not cleared during an occlusion, and it reinjects the freshest buffered frames into the model's conditioning set only on frames where the recency window is incomplete. Whenever tracking is interrupted,

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

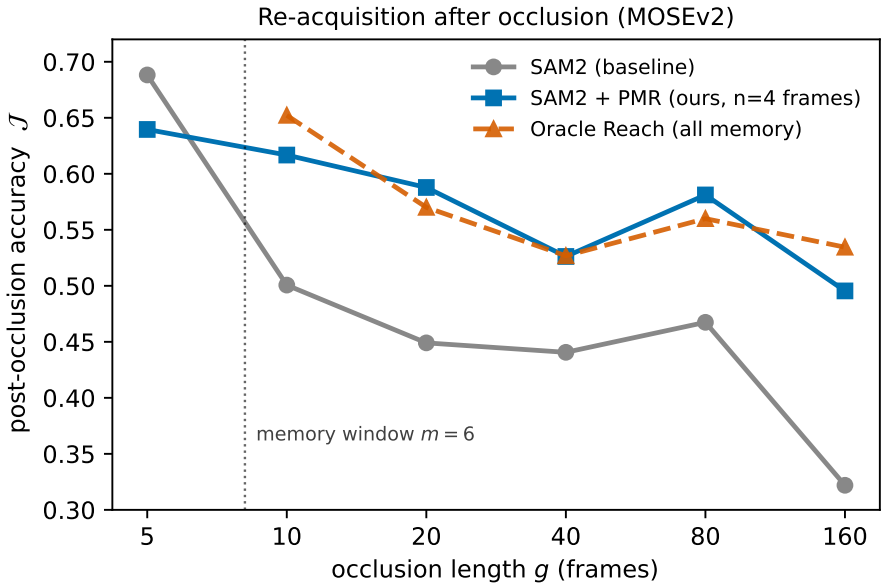


Figure 1: Post-occlusion accuracy as a function of occlusion length on MOSEv2. The SAM2 baseline collapses as soon as the occlusion exceeds its fixed memory window (m=6) and then plateaus: longer occlusions are no worse because the window is already empty and the model has fallen back on the first-frame template. Persistent Memory Reinjection (PMR), using only the n=4 freshest pre-occlusion frames, recovers a large share of the lost accuracy at every length beyond the window, and essentially matches *Oracle Reach*, which attends to all past memory without any horizon (Section 4.4).

the incompleteness of the window is a free, observable signal that coincides exactly with the interruption, so PMR is inert during normal tracking and activates exactly when the object has been out of view. In unconstrained video the model writes a memory on every frame, occluded or not, so the window never develops holes; there, the same reinjection is applied always-on or under a confidence trigger (Section 4.8).

On a controlled occlusion benchmark over 1500 MOSEv2 sequences, PMR delivers large and statistically reliable gains (Figure 1). It raises post-occlusion region accuracy ($\mathcal{J}$) by $+0.086$ to $+0.139$ for occlusions of 10 to 80 frames (and by $+0.174$ at 160), with bootstrap 95% confidence intervals that exclude zero at every length, and it raises the fraction of objects successfully re-acquired by 9 to 20 percentage points. The mechanism transfers to DAVIS-2017, where it improves medium-occlusion accuracy by up to $+0.064$, and it leaves both standard MOSEv2 and DAVIS tracking *exactly* unchanged (paired difference of 0.0000), so the gains come at no cost on clean video. Finally, fine-tuning SAM2 end-to-end improves continuous tracking but leaves post-occlusion accuracy untouched, whereas PMR's gain persists undiminished on the fine-tuned model, demonstrating that re-acquisition is bottlenecked by the memory horizon and not by the quality of the learned representation.

2 Background and Related Work

Semi-supervised VOS has progressed from per-video adaptation, through learned correspondence and external memory, to object-level and promptable foundation models. We summarize that arc and then locate the gap PMR targets: prior work has made memory more accurate, selective, or scalable, but has rarely treated *reachability under a bounded streaming context* as its own failure mode: the case where the relevant memory exists, is recent relative to the last reliable observation, and is already encoded, yet is no longer reachable because the online attention window has moved past it.

Benchmarks. DAVIS standardized dense masks and the $\mathcal{J}\&\mathcal{F}$ protocol (Perazzi et al., 2016; Pont-Tuset et al., 2017), and YouTube-VOS scaled category diversity (Xu et al., 2018), but their targets are largely salient and isolated, the regime where aggregate $\mathcal{J}\&\mathcal{F}$ saturates while realistic deployment still fails on small, crowded, occluded, or temporarily absent objects. MOSE was built to stress these conditions (Ding et al., 2023); our experiments use MOSEv2, which is larger and deliberately harder, with severe occlusion and reappearance, tiny objects, adverse conditions, and camouflage (Ding et al., 2025a). PMR targets exactly the long-horizon identity failures these complex-scene benchmarks expose, not already-saturated clean video.

From adaptation to memory-based propagation. Early methods adapted an image segmenter to the first-frame mask (Caelles et al., 2017; Voigtlaender & Leibe, 2017), then learned matching and propagation replaced per-video optimization (Voigtlaender et al., 2019; Yang et al., 2020; Luiten et al., 2018; Seong et al., 2020). Space-Time Memory networks established the standard key–value memory interface (Oh et al., 2019), refined by STCN and MiVOS (Cheng et al., 2021b;a) and extended to long videos by XMem's multi-store consolidation (Cheng & Schwing, 2022). Crowded scenes added an identity pressure addressed by object-level propagation in AOT, DeAOT, and Cutie (Yang et al., 2021; Yang & Yang, 2022; Cheng et al., 2024). This line asks which representation to store and how to read it; PMR instead asks whether the *right already-stored frames are reachable at re-acquisition*, which can fail even when memory features are good. Relatedly, memory-management work disagrees on whether to keep more or less history: XMem consolidates, MiVOS uses top-k reading, and RMem argues restricted banks help (Cheng & Schwing, 2022; Cheng et al., 2021a; Zhou et al., 2024). PMR resolves this for occlusion re-acquisition: rather than uniformly expanding or shrinking memory, it keeps a small persistent buffer outside the recency window and reinjects a few frames only when that window is incomplete, a selective-reach rather than memory-capacity mechanism.

From SAM to memory-centered video models. SAM made segmentation promptable (Kirillov et al., 2023), and the first SAM-for-video systems paired it with external propagation, point tracking, or decoupled temporal modules (Yang et al., 2023; Cheng et al., 2023b; Rajič et al., 2025; Cheng et al., 2023a). SAM2 instead folds a streaming memory into the foundation model itself (Ravi et al., 2024); its bounded recency window is precisely what creates the horizon failure we study. Recent extensions modify SAM2's memory for long video, tracking, and efficiency: SAM2Long maintains a training-free memory tree of pathways (Ding et al., 2025b), SAMURAI and DAM4SAM add motion- and distractor-aware memory selection (Yang et al., 2024; Videnovic et al., 2025), EfficientTAM and EdgeTAM compress the pipeline (Xiong et al., 2025; Zhou et al., 2025), and MedSAM-2 repurposes it with a confidence-sorted memory bank (Zhu et al., 2024). PMR is narrower than all of these: it does not branch futures, learn a memory policy, add motion, or compress attention. It simply keeps the freshest reliable pre-occlusion memories reachable when the recency window has holes.

Re-identification and bounded context. The reappearance problem also appears in multi-object tracking, from motion association in SORT to appearance embeddings in DeepSORT and robust association in ByteTrack (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2022); these import better matchers, whereas PMR first ensures the matcher even sees the most relevant pre-disappearance evidence. The horizon effect is likewise familiar in streaming sequence models, where Transformer-XL caches states and Longformer sparsifies attention to extend context (Dai et al., 2019; Beltagy et al., 2020). For re-acquisition, however, the useful long-range signal is sparse and concentrated in a few pre-occlusion frames and is needed only when normal context becomes unreliable, so PMR realizes this as sparse, gated reinjection rather than global context expansion.

3 Persistent Memory Reinjection

We first describe how SAM2 selects the memories it conditions on, make precise the failure this selection induces under occlusion, and then present PMR (Figure 2). Throughout, "memory frame" refers to the encoded representation SAM2 produces for a tracked frame (its spatial memory features together with its object-pointer token), which is exactly the object that SAM2's memory attention consumes.

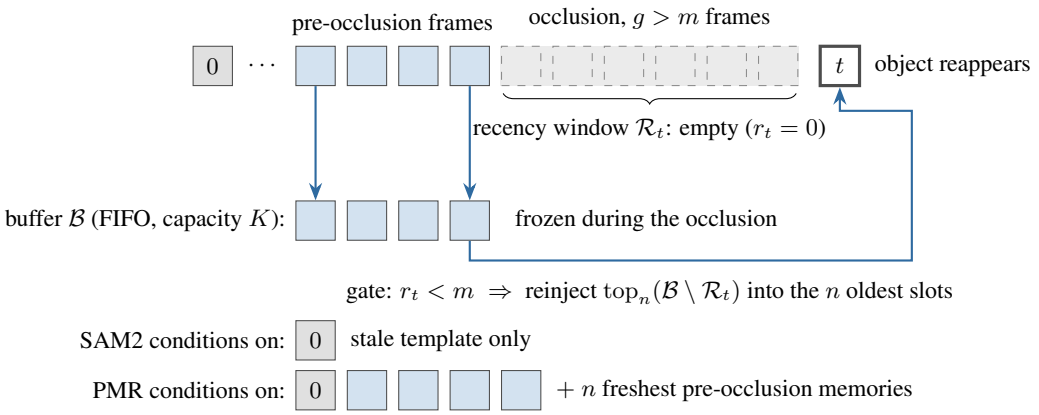


Figure 2: **PMR overview.** SAM2 conditions each frame on the first frame plus the m most recent *tracked* frames. After an occlusion longer than m, every index in the recency window falls inside the occlusion, so the model is reduced to the stale first-frame template (top). PMR maintains a small FIFO buffer $\mathcal{B}$ of the memories SAM2 already computed; because $\mathcal{B}$ is written only when a frame is tracked, it is frozen during the occlusion and still holds the freshest pre-occlusion memory when the object reappears. A structural gate compares the number of present recent frames r_t to m: when the window is incomplete, the n most recent buffered memories are re-keyed into the n oldest attention slots (displacing, not appending; Eq. 2); when it is complete, PMR is exactly SAM2. In unconstrained video, where memory is written every frame and the window never develops holes, the same reinjection runs always-on or under a confidence trigger with a conservatively admitted buffer (Section 4.8).

Recency-windowed conditioning. When SAM2 segments frame t during tracking, it conditions on a set of memory frames

$$\mathcal{M}_t \;=\; \{\,0\,\} \;\cup\; \mathcal{R}_t, \qquad \mathcal{R}_t \;=\; \{\,t-1,\, t-2,\, \ldots,\, t-m\,\} \cap \mathcal{T}_t, \tag{1}$$

where $\{0\}$ is the first (conditioning) frame, m is the fixed window size (in SAM2-Hiera-L, $m = 6$), and $\mathcal{T}_t$ is the set of frames already tracked before t. The memory features of the frames in $\mathcal{M}_t$ are concatenated and read by cross-attention. During uninterrupted tracking every recent frame has been tracked, so $\mathcal{R}_t$ is full ($|\mathcal{R}_t| = m$) and selection reduces to "the last m frames."

The occlusion failure. Suppose the object is occluded over an interval of length g ending at frame t_e, during which no reliable memory is written. For the first frames after reappearance, the indices $t-1, \ldots, t-m$ fall inside the occlusion and were never tracked, so $\mathcal{R}_t$ is empty and $\mathcal{M}_t = \{0\}$: the model conditions on the first-frame template alone. Whenever $g > m$, *every* pre-occlusion memory frame has already left the window. The first-frame template is the appearance the object had at $t = 0$, which after a long video is typically a poor match for the reappearing object, so the model fails to recover the correct identity and the failure cascades. The relevant memory is the last frames before the occlusion, whose appearance is closest to the reappearing object, and it exists in the model's history; recency selection simply cannot reach it once $g > m$.

Persistent buffer. PMR adds one piece of state: a first-in-first-out buffer $\mathcal{B}$ of capacity K (we use $K = 16$). After each tracked frame, its encoded memory is appended to $\mathcal{B}$ provided the predicted mask is non-empty; when more than K entries accumulate, the oldest is evicted. This is the only admission rule in every controlled-occlusion-benchmark experiment; a confidence-filtered (*conservative*) variant of the same buffer, used for natural failures where the model keeps writing unreliable memory, is defined in Section 4.8. Two properties follow. First, $\mathcal{B}$ is written only when a frame is tracked, so during an occlusion in which the model receives no observations nothing is evicted: the buffer retains the K freshest pre-occlusion frames for the entire occlusion, *regardless of its length*. K bounds how many pre-occlusion frames are retained, not how long an occlusion can be bridged. Second, unlike the attention window, $\mathcal{B}$ is not indexed by absolute recency relative to t, so its entries remain addressable after the window has moved past them. Maintaining $\mathcal{B}$ costs K stored memory frames and no extra computation on any frame.

Gated reinjection. PMR changes the conditioning set only when the recency window is incomplete, which is the observable signature of an interrupted track. Concretely, on frame t we measure the number of present recent frames $r_t = |\mathcal{R}_t|$. If $r_t = m$ the window is intact, we do nothing, and PMR is exactly equivalent to SAM2. If $r_t < m$ there are holes, and we *overwrite* the n oldest of the m attention slots (positions $t{-}m, \ldots, t{-}(m{-}n{+}1)$) with the n most recent entries of $\mathcal{B}$ that are not already in $\mathcal{R}_t$, keeping any recent frames that survive in the $m{-}n$ newest positions:

$$\widehat{\mathcal{M}}_t = \begin{cases} \mathcal{M}_t, & r_t = m \\ \{0\} \cup \mathcal{R}_t^{\text{new}} \cup \text{top}_n(\mathcal{B} \setminus \mathcal{R}_t), & r_t < m, \end{cases} \tag{2}$$

where $\mathcal{R}_t^{\text{new}} = \mathcal{R}_t \cap \{t{-}1, \ldots, t{-}(m{-}n)\}$ are the surviving recent frames in the $m{-}n$ newest positions. Reinjection therefore *displaces* rather than appends: $|\widehat{\mathcal{M}}_t| \leq m + 1$ always, matching SAM2's native slot count (e.g., with $r_t = 5$ and $n = 4$, the reinjected entries replace the surviving recent frames in the four oldest positions, and the two newest recent frames are kept). The reinjected frames take the oldest temporal positions, consistent with their being older than any surviving recent frame. The model's cross-attention is otherwise untouched: it performs the same localization it always does, now over a conditioning set that includes the freshest pre-occlusion memory. The gate has two consequences. First, PMR provably does not affect continuous tracking, because on every frame with a full window it returns $\mathcal{M}_t$ unchanged. Second, PMR activates exactly on the frames where the object has been out of view, which is precisely the regime in which it is needed.

Why reinjection is sufficient. PMR uses no identity embedding, learned retrieval, reliability classifier, or separate recovery action. It changes not *what* any memory says but only how far back reachable memory extends, and only during occlusion. Section 4.7 shows empirically that the machinery one might expect to need (selecting diverse past appearances, learning which memory to retrieve, or re-detecting the object) does not improve over reinjecting the most recent pre-occlusion frames: the failure is one of *reach*, not content or selection, so once the freshest pre-occlusion memory is reachable, SAM2's existing attention suffices to localize the object.

Implementation. PMR is implemented entirely at inference: we never modify SAM2's memory-attention module but expose the buffered frames to it by editing the per-object memory dictionary it already reads, re-keying the reinjected entries onto the temporal slots the attention expects (which keeps the object-pointer and positional-encoding pathways intact, with a separate buffer per object; full details in Appendix A). We release code as a drop-in subclass of the SAM2 video predictor.

4 Experiments

4.1 Setup

Data and metrics. We evaluate on MOSEv2 (Ding et al., 2025a), built around crowded scenes and frequent occlusions, and on DAVIS-2017 (Pont-Tuset et al., 2017) for generalization, scoring region similarity $\mathcal{J}$ (mask IoU), boundary accuracy $\mathcal{F}$, and their $\mathcal{J}\&\mathcal{F}$ mean. As the MOSEv2 validation set releases only first-frame ground truth, all MOSEv2 results use held-out training sequences with dense annotations. Continuous-tracking experiments (Section 4.6; the continuous rows of Sections 4.9, 4.10) use sequence indices [200:700] (500 sequences), the natural-failure experiments (Sections 4.8, 4.9) use [200:1000] (800), and the controlled occlusion benchmark uses [200:1700] (1500) to retain adequate sample sizes at long occlusions. All are disjoint from the sequences used to train auxiliary components ([0:200]) or to fine-tune ([1700:3200]). The base model is SAM2.1-Hiera-L; PMR adds no parameters.

Occlusion benchmark. To measure re-acquisition directly we use a controlled occlusion protocol. For each sequence we choose an occlusion window of length g starting at 30% of the video; we track up to the window, skip the window entirely so the model receives no observations inside it (memory and the PMR buffer persist), and resume tracking afterward. We report the mean post-occlusion $\mathcal{J}$ and the re-acquisition rate (fraction with post-occlusion $\mathcal{J} > 0.5$) as functions of g, with paired differences against the SAM2 baseline and bootstrap 95% confidence intervals over sequences (10,000 resamples; significant when the interval excludes zero). The unit of analysis is the *sequence*: all annotated objects are tracked, and the first annotated object of each sequence is

Table 1: Re-acquisition after a controlled occlusion of length g (MOSEv2, post-occlusion $\mathcal{J}$). The baseline collapses as soon as the occlusion exceeds the memory window (m=6), and does not recover with longer occlusions. PMR restores a large fraction of the lost accuracy at every length beyond the window. "Recov." is the fraction of objects re-acquired ($\mathcal{J} > 0.5$). All PMR margins are bootstrap-significant. n is the number of qualifying sequences.

	Baseline		PMR (ours)			
g (frames)	$\mathcal{J}$	Recov.	$\mathcal{J}$	$\Delta\mathcal{J}$ (95% CI)	Recov.	n
10	0.501	55%	**0.617**	$+0.116$ $[+0.094, +0.139]$	68%	730
20	0.449	49%	**0.588**	$+0.139$ $[+0.112, +0.167]$	64%	574
40	0.441	49%	**0.526**	$+0.086$ $[+0.054, +0.118]$	58%	339
80	0.467	49%	**0.581**	$+0.114$ $[+0.054, +0.175]$	63%	112
160	0.322	32%	**0.495**	$+0.174$ $[+0.057, +0.307]$	52%	31

scored. A sequence qualifies at occlusion length g if the video is long enough to hold the window ($\geq g+25$ frames, with at least 10 frames remaining after it), the target mask covers at least 100 pixels immediately before and after the window, and at least three post-window frames are scoreable; n in each table cell is the number of qualifying sequences, which decreases with g.

4.2 The memory horizon limits re-acquisition

Our first experiment establishes the diagnosis. The "Baseline" columns of Table 1 show that as g crosses the window size m=6, post-occlusion accuracy drops sharply and then *stops depending on* g: it is 0.501 at g=10 and a statistically indistinguishable 0.44 to 0.47 from g=20 through g=80 (the flat gray curve in Figure 1). Once the window is empty no further memory remains to be lost, so longer occlusions do not worsen conditioning; the model is already reduced to the first-frame template. A drift account would instead predict smooth monotonic worsening; the plateau is direct evidence that a large part of the failure is a step loss of *reach* at g=m, not gradual corruption.

4.3 Persistent memory recovers long-occlusion objects

Reinjecting persisted pre-occlusion memory recovers a large share of the lost accuracy. The "PMR" columns of Table 1 (default K=16, n=4) improve post-occlusion $\mathcal{J}$ by $+0.116$, $+0.139$, $+0.086$, $+0.114$, and $+0.174$ at $g = 10, 20, 40, 80, 160$, every margin bootstrap-significant, and raise the re-acquisition rate by 9 to 20 points. The gain does not decay with occlusion length; if anything it is largest at g=160, where the baseline degraded most. Because the buffer is written only on tracked frames, it is frozen for the entire occlusion and the freshest pre-occlusion frames remain buffered and equally useful regardless of duration, including $g \gg K$ (examined below). The most damaging VOS failures, permanent loss after a long occlusion, are thus substantially mitigated by making a few already-computed memory frames reachable at re-acquisition, with no training or architectural change.

4.4 How much signal PMR captures and how much memory it needs

We next assess whether PMR leaves recoverable signal unexploited: a small reinjection might help while full access to history would help substantially more. We do this with an infinite-horizon ceiling. *Oracle Reach* removes the memory horizon altogether by enlarging the attention window to encompass every past frame (implemented with a window of N=256, longer than any evaluated video, and an extended temporal-position embedding; see Appendix B). It is an upper bound on what recency-free access to memory can provide. Table 2 compares it to PMR.

Oracle Reach is a stronger tracker: on the 500-sequence continuous-tracking set it improves $\mathcal{J}$ by $+0.011$ (paired 95% CI $[+0.003, +0.020]$; $+0.010$ $\mathcal{J}\&\mathcal{F}$; $+0.030$ $\mathcal{J}$ on failure-episode sequences; full breakdown in Table A2), showing that conditioning on more reachable history is helpful. Despite Oracle Reach having access to every past frame, PMR with a four-frame buffer matches it to within 0.04 $\mathcal{J}$ at every occlusion length and equals its mean gain ($+0.126$ versus $+0.133$). The recoverable signal is therefore highly concentrated: the freshest few pre-occlusion frames carry almost all of it,

Table 2: PMR versus an infinite-horizon oracle and across buffer sizes (MOSEv2, $\Delta\mathcal{J}$ over baseline; all entries bootstrap-significant). *Oracle Reach* removes the horizon entirely and attends to all past memory. PMR, using only the four freshest pre-occlusion frames, matches it to within 0.04 at every length. Buffer size K is irrelevant: $K{=}16$, 32, and 64 are identical to three decimals, and the gain does not collapse when the occlusion exceeds the buffer ($g{>}K$).

g	PMR ($K{=}16$)	Oracle Reach	PMR $K{=}32$	PMR $K{=}64$	baseline $\mathcal{J}$
10	+0.116	**+0.152**	+0.116	+0.116	0.501
20	+0.139	+0.121	+0.138	+0.138	0.449
40	+0.086	+0.086	+0.085	+0.085	0.441
80	+0.114	+0.093	+0.114	+0.114	0.467
160	+0.174	**+0.213**	+0.174	+0.174	0.322
mean	+0.126	+0.133	+0.125	+0.125	

Table 3: Re-acquisition on DAVIS-2017 (post-occlusion $\mathcal{J}$). PMR transfers to a different dataset, with significant gains on medium occlusions and no harm on short ones. DAVIS has fewer and shorter occlusions than MOSEv2, so absolute gains are smaller.

g (frames)	Baseline $\mathcal{J}$	PMR $\mathcal{J}$	$\Delta\mathcal{J}$ (95% CI)	n
10	0.856	0.880	$+0.024$ $[+0.000, +0.056]$	87
20	0.787	0.851	$\mathbf{+0.064}$ $[+0.017, +0.117]$	79
40	0.814	0.834	$+0.020$ $[-0.001, +0.056]$	53

and an unbounded memory adds little beyond them, so a minimal reinjection recovers nearly all of what removing the horizon would buy.

Removing the horizon does not repair re-acquisition fully. Pre-occlusion accuracy is about 0.80 $\mathcal{J}$; the occlusion drops the baseline to 0.32–0.50, and both PMR and Oracle Reach recover only a quarter to a half of that loss. The bounded horizon is thus a substantial, training-free-fixable contributor but not the majority cause; most of the loss persists even with unlimited memory, presumably because the object's appearance has changed beyond what any stored frame captures. What the oracle does establish is that *of the accuracy recoverable through memory at all*, a bounded horizon is the main limiting factor and a four-frame buffer captures nearly all of it.

Finally, the same table shows PMR needs only a few frames, not a larger horizon. Buffer size has no measurable effect ($K{=}16$, 32, 64 are identical to three decimals), because PMR conditions on only the $n{=}4$ freshest entries; and the gain does not collapse when the occlusion outlasts the buffer: at $g{=}160 \gg K{=}16$ it is largest ($+0.174$), because in a clean occlusion the buffer is not overwritten while the object is absent. PMR is thus a sparse, gated reinjection of the few frames that carry the signal, not a wider window, with negligible overhead.

4.5 Generalization to DAVIS-2017

Running the same protocol on DAVIS-2017 (Table 3), PMR again improves post-occlusion accuracy, significantly at $g{=}20$ ($+0.064$) and $g{=}10$ ($+0.024$). Gains are smaller than on MOSEv2 because DAVIS has far fewer long occlusions (2% of sequences fail under continuous evaluation vs. 41% of MOSEv2), so there is less re-acquisition to do; the mechanism transfers without dataset-specific tuning.

4.6 Persistent memory is free on clean video

A re-acquisition method must not degrade ordinary tracking. Because PMR's gate returns the conditioning set unchanged whenever the recency window is full (always the case during uninterrupted tracking), PMR is provably inert on clean video. Table 4 confirms this: on both MOSEv2 and DAVIS, PMR matches the baseline $\mathcal{J}$ and $\mathcal{J}\&\mathcal{F}$ to four decimals (paired per-sequence difference exactly 0.0000), including on sequences that contain a failure episode. The re-acquisition benefit thus comes at no cost where it is not needed. We emphasize the scope of this guarantee: it certifies

Table 4: Continuous tracking (no induced occlusion). PMR's gate keeps it exactly inert when the memory window is intact, so it matches the baseline to four decimals on both datasets. "Fail. J" is mean $\mathcal{J}$ on the subset of sequences that contain a failure episode.

	$\mathcal{J}$	$\mathcal{J}\&\mathcal{F}$	Fail. $\mathcal{J}$
MOSEv2 baseline	0.7750	0.6995	0.6288
MOSEv2 + PMR	0.7750	0.6995	0.6288
DAVIS baseline	0.8959	0.8317	0.7375
DAVIS + PMR	0.8959	0.8317	0.7375

Table 5: Design analysis (MOSEv2 occlusion benchmark, $\Delta\mathcal{J}$ vs baseline). Reinjecting the most recent pre-occlusion frames is what matters. The number of reinjected slots n has a mild optimum at 4; buffer size beyond $K{=}16$ does not help; and neither learned identity retrieval, an oracle that selects buffered frames by ground-truth overlap, nor a separate point-prompt recovery action improves over plain recency reinjection. Bold cells are bootstrap-significant.

Variant	$g{=}10$	$g{=}20$	$g{=}40$	$g{=}80$
PMR, $n{=}4$ (default)	**+0.116**	**+0.139**	**+0.086**	**+0.114**
$n{=}2$	+0.111	+0.121	+0.071	+0.107
$n{=}6$	+0.096	+0.103	+0.076	+0.097
buffer $K{=}32$	+0.116	+0.138	+0.085	+0.114
Selection of buffered frames:				
learned identity retrieval	+0.008	+0.012	+0.009	+0.016
oracle (by GT overlap)	+0.007	+0.015	+0.014	+0.018
Alternative action: point-prompt recovery	+0.021	+0.015	+0.014	+0.003

the *gated* variant, whose gate by construction never fires on uninterrupted video. The always-on variant used for natural failures (Section 4.8) modifies the conditioning set on every frame, so its clean-video cost is not covered by this argument and must be measured directly: on the same 500-sequence set it scores 0.7744 $\mathcal{J}$ (0.6991 $\mathcal{J}\&\mathcal{F}$), a paired difference from the baseline of -0.0008 $\mathcal{J}$ (95% CI $[-0.009, +0.007]$) and -0.0006 $\mathcal{J}\&\mathcal{F}$ ($[-0.008, +0.007]$), i.e. no measurable cost even though it alters the conditioning set on 84% of sequences.

4.7 What matters in the design

We vary one component at a time on the occlusion benchmark to identify which parts of the design account for the gain (Table 5). The results indicate that the mechanism depends on reach rather than content or selection.

First, the gain is governed by how many recent pre-occlusion frames are reinjected, with a mild optimum: $n{=}4$ slightly exceeds $n{=}2$ and $n{=}6$ (filling all six slots displaces the few genuine post-occlusion frames that do exist). Enlarging the buffer beyond what an occlusion spans has no effect ($K{=}16$ and $K{=}32$ are identical), confirming that only the freshest pre-occlusion frames carry the signal.

Second, the most recent buffered frames perform best. A natural alternative is to select a diverse or learned-best subset of past appearances under the intuition that the right re-identification cue may lie further back. We test this by replacing recency selection with a learned identity-similarity retrieval, and separately with an oracle that selects buffered frames by their ground-truth overlap with the target at the query frame; both operate over a buffer whose admission favors *diverse* past appearances (rather than the pure recency FIFO), since that is the bank such selection strategies presuppose (exact configurations in Appendix B). Learned retrieval yields essentially no gain, and even the oracle selection underperforms plain recency: the diverse entries it can choose from are necessarily older, and (consistent with the reach-not-content diagnosis) older appearances carry little conditioning signal however well they are chosen. For *conditioning* (as opposed to retrieval-based

Table 6: Natural-failure recovery by category (MOSEv2).

Failure type	share	baseline $\mathcal{J}$	always	confidence
Scale change	43%	.025	+.134	+.097
Occlusion	21%	.021	+.142	+.086
Collapse	15%	.069	+.117	+.050
Distractor	3%	.045	+.189	+.005
Other	16%	.006	+.271	+.250
All	100%	.029	+.156	+.108

matching), the frame whose appearance is closest to the reappearing object is the most recent one before the occlusion, so recency is a strong prior for conditioning.

Third, restoring reachable memory is sufficient in this benchmark; no separate recovery action is needed. We compare against augmenting the tracker with a point-prompt recovery step that, when an internal reliability detector fires, generates candidate locations and re-prompts the segmenter. This action helps only marginally on the occlusion benchmark and is dominated by reinjection, and (as detailed in Appendix B) it degrades continuous tracking because the detector occasionally fires on normal frames. PMR avoids this failure mode in the controlled setting: its trigger is the structural incompleteness of the window rather than a learned signal.

4.8 Natural failures also benefit

The mechanism also helps failures that arise during ordinary continuous tracking, where the recency window remains full and the structural gate cannot fire. Using the conservative buffer described in Appendix B, we evaluate always-on and confidence-triggered reinjection on 700 failure episodes. The always-on variant improves episode-frame accuracy by $+0.156$ $\mathcal{J}$ on average, while the confidence trigger improves it by $+0.108$; scale changes, occlusions, distractor takeovers, and segmentation collapses all benefit. The full taxonomy and statistical details are in Table A3, with qualitative examples in Figure A2. For unconstrained video, the open problem is designing a trigger that is robust to model overconfidence.

4.9 Comparison with SAM2Long

SAM2Long is the closest training-free comparison: it manages multiple memory pathways, whereas PMR changes which already-computed memories are reachable. On identical-input continuous and natural- failure evaluations, the methods are comparable within the m=6 window, but PMR significantly outperforms SAM2Long once the target is absent beyond the window ($+0.072$, bootstrap 95% CI $[+0.004, +0.151]$). PMR is also cost-free on continuous tracking, while SAM2Long uses multiple pathways per object. The full comparison and protocol are in Appendix C and Tables A4–A5.

4.10 The gain is orthogonal to fine-tuning

Fine-tuning cannot substitute for PMR: it improves continuous tracking by $+0.023$ $\mathcal{J}\&\mathcal{F}$ but leaves post-occlusion accuracy statistically unchanged, because it cannot change the fixed recency horizon. PMR's gain remains $+0.09$ to $+0.14$ $\mathcal{J}$ on the fine-tuned model and stays inert on clean video, so the two improvements are complementary. Full paired results are in Appendix D and Table A6.

5 Conclusion

A substantial component of long-occlusion failure is the fixed temporal horizon of recency-selected memory. PMR keeps a small pre-occlusion buffer reachable and reinjects it when the window is incomplete, recovering up to 17 $\mathcal{J}$ points on MOSEv2 and 6 on DAVIS while leaving continuous tracking unchanged. A four-frame buffer matches an infinite-horizon oracle, and the same mechanism helps several natural failure types, showing that reachability—not memory content or capacity—is the main limiting factor. Appendix E confirms transfer across the SAM2 family (Tables A7–A11); trigger design remains open for unconstrained video, where failures do not create observable window holes.

Reproducibility Statement

The appendix describes the inference-time integration, buffer and gating rules, evaluation protocols, statistical tests, and all ablations. The accompanying supplementary code archive contains the PMR evaluation scripts and configuration files needed to reproduce the reported comparisons, subject to the publicly released SAM2 checkpoints and datasets.

AI Use Statement

In this work, we have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning or reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. The rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, figures, and supplementary artifacts.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing (ICIP)*, 2016.

Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson–shiffrin memory model. In *European Conference on Computer Vision (ECCV)*, 2022.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander G. Schwing. Putting the object back into video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv:2305.06558*, 2023b.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H. S. Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip H. S. Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv:2508.05630*, 2025a.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. SAM2Long: Enhancing SAM 2 for long video segmentation with a training-free memory tree. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision (ACCV)*, 2018.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv:1704.00675*, 2017.

Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024.

Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *European Conference on Computer Vision (ECCV)*, 2020.

Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with SAM2. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *British Machine Vision Conference (BMVC)*, 2017.

Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. FEELVOS: Fast end-to-end embedding learning for video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing (ICIP)*, 2017.

Yunyang Xiong, Chong Zhou, Xiaoyu Xiang, Lemeng Wu, Chenchen Zhu, Zechun Liu, Saksham Suri, Balakrishnan Varadarajan, Ramya Akula, Forrest Iandola, Raghuraman Krishnamoorthi, Bilge Soran, and Vikas Chandra. Efficient track anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. YouTube-VOS: A large-scale video object segmentation benchmark. In *European Conference on Computer Vision (ECCV)*, 2018.

Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. SAMURAI: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv:2411.11922*, 2024.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv:2304.11968*, 2023.

Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision (ECCV)*, 2020.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision (ECCV)*, 2022.

Chong Zhou, Chenchen Zhu, Yunyang Xiong, Saksham Suri, Fanyi Xiao, Lemeng Wu, Raghuraman Krishnamoorthi, Bo Dai, Chen Change Loy, Vikas Chandra, and Bilge Soran. EdgeTAM: On-device track anything model. *arXiv:2501.07256*, 2025.

Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. RMem: Restricted memory banks improve video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical SAM 2: Segment medical images as video via segment anything model 2. *arXiv:2408.00874*, 2024.