TRUST, BUT VERIFY TEMPORAL MEMORY: PROCEDURE-CONSISTENCY FIREWALL FOR SURGICAL VIDEO SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal memory provides valuable context for online surgical video segmentation, but most existing methods assume that the previous states remain relavant to the current scene. We find that this assumption can break under procedural context shifts. When the query memory is obtained from an incompatible procedural context, harmful temporal information can corrupt current predictions and damage future states. Estimating memory utility with a generic indicator of does not fully resolve this problem, as such gate tends to be overly conservative and rejects useful history. Motivated by these observations, we propose a **Procedure-Consistency Memory Firewall** and a **Trustworthy Memory Writer**. The Procedure-Consistency Memory Firewall is a annotation-free, inference-time mechanism that compares the current frame procedure distribution with the provenance distribution attached to temporal memory. The Trustworthy Memory Writer determines whether the resulting state is sufficiently reliable to update trusted memory, limiting dirty context from propagating forward. Across a controlled suite of temporal interventions, our firewall selectively targets provenance mismatch rather than broadly suppressing temporal context. Our method clean performance under clean conditions while improving robustness under core stress by 6.31% and worst-case Dice by 32.77%, with the largest gain under cross-procedure mismatch. These results suggest that robust temporal reasoning depends not only on remembering the past, but also on verifying whether the memory is suitable before reuse and propagation.

1 INTRODUCTION

Temporal context is a central source of strength in video segmentation. Recent methods propagate object tokens, pixel features, or compressed states across frames, allowing evidence from an uncomplicated observation to disambiguate a difficult one later in the sequence [5, 7, 8, 39, 50]. This advantage is particularly attractive in surgical videos, where smoke, blood, occlusion, motion blur, and sudden changes of perspective often obscure the current frame [10, 29, 30]. However, persistent memory also creates a failure surface that frame-wise evaluation does not expose: a state written under one semantic context can be read after the video has transitioned to another procedure, phase, or scene. The retrieved features may remain clear and individually plausible while being wrong for the current context. Once consumed, this dreadful evidence may interfere with the predictions, and be written back to memory, turning the local mismatch into persistent errors. In other words, temporal memory is not only a source of information but also a stateful channel through which errors can propagate.

Existing video segmentation methods largely focus on *what* to remember, *where* to retrieve it, and *how* to compress or update it. Reliability mechanisms likewise tend to calculate the apparent quality, confidence, or utility of an individual memory item [16, 24, 37]. These criteria do not answer a distinct question: *does this memory belong to the same semantic context as the current observation?* High-quality features of incompatible procedures can be more harmful than noisy features of the correct ones. Conversely, the suppression of all uncertain history sacrifices the temporal evidence required under visual degradation. Therefore, we argue that robust temporal segmentation depends on distinguishing two axes that are commonly confused: the intrinsic quality of a memory item and its provenance compatibility with the current frame.

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

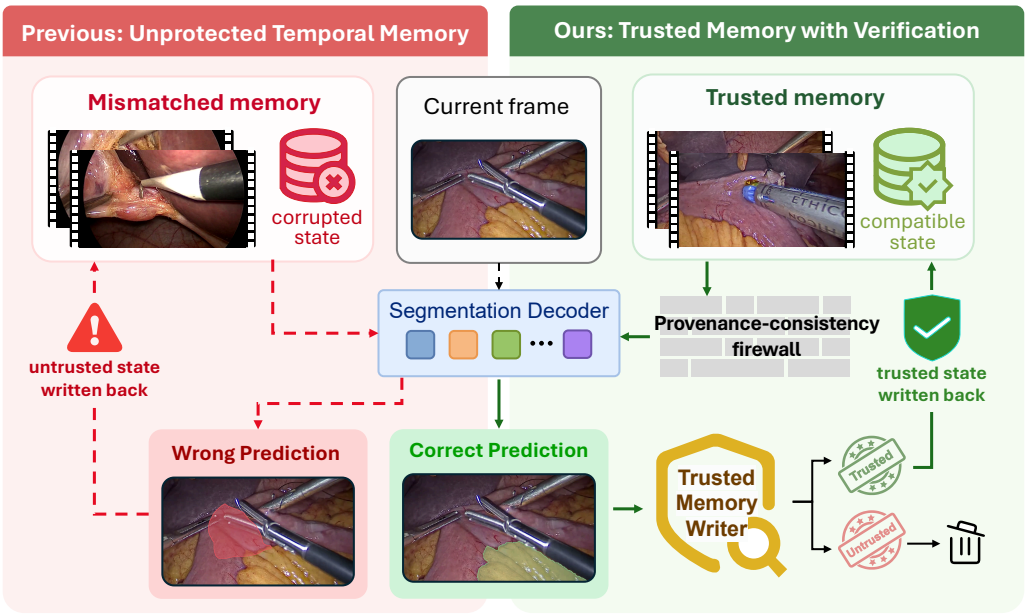

Figure 1: **Motivation.** Temporal memory is useful but should not be trusted unconditionally. Unprotected memory can propagate incompatible states and recursively contaminate future predictions. We protect memory reads with a provenance consistency firewall and control memory updates with a trusted memory writer.

We study this distinction through controlled memory interventions rather than through aggregate clean-video accuracy alone. Our protocol retains the current frame and segmentation target while changing the origin and degradation of the supplied history, including matched history, visually corrupted history, and cross-procedure history. In ATLAS's query memory model and two additional temporal models, ConvGRU and TMANet, the most damaging cases are not always those with the lowest visual quality. Semantically incompatible but otherwise plausible history can produce large, persistent failures. This observation also reveals why the general memory utility gate is insufficient. It may reject useful context affected by corruption while accepting clean memory from an incompatible procedure. The memory read decision should therefore be based on comparing the provenance of the current scene with that of the retrieved memory.

Inspired by this observation, we propose the *Procedure-Compatibility Memory Firewall* (PCMF), a lightweight control layer for temporal segmentation. At each memory read, PCMF estimates procedure distributions for the current scene and the candidate memory context and compares their difference using Jensen-Shannon divergence [28]. Compatible evidence follows the model's original temporal path, while sufficiently incompatible evidence is rejected. The read firewall is coupled with a *Trusted Memory Writer* so that rejected context is not immediately reintroduced through the next recurrent state. Notably, the method uses predicted semantic evidence rather than externally supplied procedure annotation at inference time. This approach adds no trainable parameters to the firewall itself and can be attached to explicit memory retrieval and recurrent states.

On the ATLAS dataset, applying PCMF to the native query-memory model improves cross-procedure global Dice from 0.2153 to 0.2859 ($+0.0706$). A paired video-level bootstrap gives a mean-frame Dice gain of 0.0639, with a 95% confidence interval of $[0.0463, 0.0855]$, while the paired clean effect remains near zero ($+0.00044$). The same principle transfers beyond the ATLAS native model. On ATLAS, TMANet gains 0.0184 global Dice under cross-procedure history, with a paired video-level gain of 0.0132 (95% confidence interval of $[0.0084, 0.0182]$), and essentially no change on paired clean inputs. On the CholecSeg8k dataset, formal tests show positive cross-procedure gains for both ConvGRU ($+0.0742$) and TMANet ($+0.0073$), while their paired clean outputs are unchanged. These results support that PCMF preserves the benefit of temporal context when it is valid and substantially limits negative transfer when it is not.

Our contributions are threefold:

- We identify *memory provenance compatibility* as a failure axis of stateful video segmentation that is distinct from memory quality and introduce controlled interventions that isolate its effect while holding the current-frame task fixed.
- We propose PCMF, a lightweight control layer that detects semantic mismatch from predicted evidence, blocks harmful cross-context transfer, and requires no external procedure annotation at inference time.
- We introduce a TRUSTED MEMORY WRITER that separates the latest sensory state from readable trusted memory and maintains a trusted state memory. With safe updates, it prevents rejected context from being written back into the next recurrent memory state.

2 RELATED WORKS

Memory-based video segmentation. Video object segmentation has progressed from per-frame adaptation and recurrence to explicit mechanisms for propagating visual evidence through video [2, 6, 38, 44, 46, 48]. Space-Time Memory Networks (STM) retrieve dense correspondences between the current frame and a bank of previously segmented frames [36]. STCN improves the coverage and efficiency of this matching [7, 32, 40, 49], while AOT propagates multiple objects through identity-aware transformer queries [50]. XMem organizes state into sensory, working, and long-term stores to support long sequences [5, 11, 12, 17, 51], and Cutie reduces noisy pixel matching through an explicit object-level representation [8, 26, 45]. More recently, Segment Anything and SAM 2 extended promptable segmentation to images and streaming video, followed by concept-driven SAM 3 [3, 22, 39]. These methods principally improve how much history is retained and how relevant content is retrieved. Our concern is distinct: even a compact, high-quality memory can be harmful when its semantic origin is incompatible with the current scene. PCMF therefore audits the provenance of the state that is actually readable before allowing it to affect the current prediction.

Temporal semantic segmentation. Video semantic segmentation exploits temporal evidence without requiring a target mask in the first frame [13, 19, 20, 23, 25, 27, 31, 35, 52]. TMANet uses temporal memory attention [47] to retrieve long-range features from past frames, while recurrent feature fusion summarizes temporal context in a compact hidden state. In surgery, Grammatikopoulou et al. attach a spatio-temporal decoder to frame encoders to improve accuracy and temporal consistency [15], while DMNet combines local recurrent state with a global memory populated according to uncertainty and frame similarity [47]. ATLAS propagates mask queries on top of foundation representations and jointly predicts surgical procedure identity [4, 10, 21, 42]. Rather than replacing these fusion mechanisms, we treat them as temporal-memory consumers and apply the same provenance-control interface around query propagation, ConvGRU states, and temporal attention. This separation allows us to evaluate whether the proposed principle generalizes across different mechanisms for representing temporal history.

Surgical scene segmentation. Surgical segmentation must handle deformable anatomy, specularities, smoke, occlusion, rare structures, and long workflow-dependent appearance changes [1, 9, 14, 33, 34, 41]. CholecSeg8k provides dense semantic annotations for laparoscopic cholecystectomy [18], whereas the multi-procedure ATLAS benchmark exposes substantially broader anatomical and procedural variation [10]. Surgical SAM 2 accelerates inference by pruning frames [29], and ReSurgSAM2 targets credible initialization and long-term tracking [30]. Existing surgical temporal models mainly seek higher accuracy or smoother predictions within an ordinarily coherent sequence. We instead use controlled history interventions to isolate a stateful failure mode: the current image can remain unchanged while the model receives plausible memory from a different procedure. This setting is a mechanism test of negative transfer, not a claim that such swaps describe their frequency in routine clinical deployment.

3 METHODOLOGY

3.1 PROBLEM SETUP AND OVERVIEW

We consider online semantic segmentation, where a model processes frames $x_1, \ldots, x_t$ sequentially and predicts a dense label map $\hat{y}_t$ without access to future frames. Let M_{t-1} denote the temporal

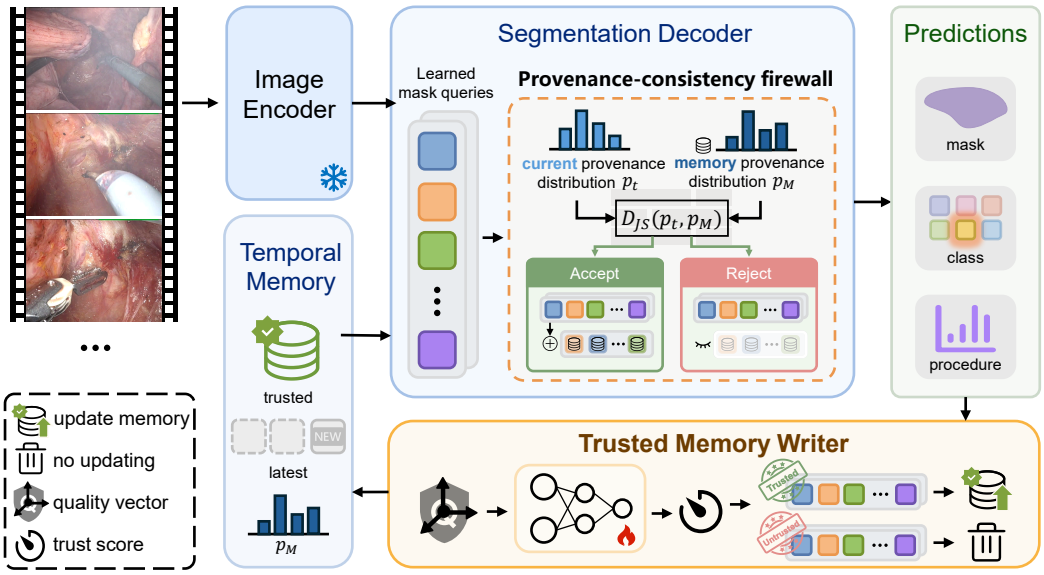

Figure 2: **Overview of our framework.** The provenance consistency firewall compares the current and memory provenance distributions using Jensen–Shannon divergence to determine whether temporal memory should be used for the current frame. The trusted memory writer then evaluates the decoded state and updates trusted memory only when its calibrated quality score is sufficiently high.

state available before processing x_t. A generic temporal segmenter can be written as

$$(\ell_t^T, Q_t^T) = F_\theta(x_t; M_{t-1}), \tag{1}$$

where ℓ_t^T are semantic logits and Q_t^T is the state candidate produced by the temporal pathway. With temporal history disabled, the same model produces an image-only anchor:

$$(\ell_t^A, Q_t^A, \mathbf{p}_t) = F_\theta(x_t; \varnothing). \tag{2}$$

Here, $\mathbf{p}_t \in \Delta^{K-1}$ denotes predicted distribution over a semantic context variable c_t, defined as surgical procedure for ATLAS and surgical phase for CholecSeg8k. Importantly, $\mathbf{p}_t$ is inferred without access to temporal history, ensuring that the state being evaluated cannot influence the evidence used to assess it.

Our objective is neither to replace the temporal model nor to suppress history temporal by default. Instead, we introduce two complementary controls around its memory read-write interface. The *Procedure-Compatibility Memory Firewall* (PCMF) determines whether the readable state is semantically compatible with the current scene, whereas the *Trusted Memory Writer* determines whether the routed state is reliable enough to be written back to memory. The former performs a relational compatibility test between the current and stored contexts, while the latter applies an intrinsic reliability criterion to the newly decoded state.

3.2 Provenance-Carrying Temporal Memory

A conventional recurrent system maintains a temporal state without explicitly tracking its semantic provenance. We instead represent readable memory as

$$\mathcal{M}_{t-1} = (M_{t-1}, \mathbf{p}_M, a_{t-1}), \tag{3}$$

where $\mathbf{p}_M$ denotes the stored semantic context distribution associated with M_{t-1} at the time of its admission to memory, and a_{t-1} denotes the age of the stored state. Provenance therefore follows the readable state, not the most recently observed frame.

To maintain this distinction, we track two separate state variables. A *sensory state* records the latest decoded state candidate and is used only for diagnostic purposes. In contrast, a *trusted state* is

updated only when a candidate passes the writer criterion and is used for formal memory reads. Whenever a new trusted state is accepted, its associated context distribution is updated atomically:

$$M_t \leftarrow Q_t \quad \Longleftrightarrow \quad \mathbf{p}_M \leftarrow \mathbf{p}_t. \tag{4}$$

If the write is rejected, both M_t and $\mathbf{p}_M$ remain unchanged, while the memory age is incremented. This atomic update ensures that a newly inferred context distribution cannot be incorrectly associated with an older readable state. At the start of a stream, or whenever no valid trusted state is available, the model uses the image anchor and bypasses the firewall.

3.3 Procedure-Compatibility Memory Firewall

Given the current predicted distribution $\mathbf{p}_t$ and stored context distribution $\mathbf{p}_M$ associated with the readable memory, we measure their discrepancy using the Jensen-Shannon divergence [28]. Defining $\mathbf{m}_t = \frac{1}{2}(\mathbf{p}_t + \mathbf{p}_M)$, we compute

$$d_t = D_{\mathrm{JS}}(\mathbf{p}_t, \mathbf{p}_M) = \tfrac{1}{2} D_{\mathrm{KL}}(\mathbf{p}_t \| \mathbf{m}_t) + \tfrac{1}{2} D_{\mathrm{KL}}(\mathbf{p}_M \| \mathbf{m}_t). \tag{5}$$

Unlike an argmax comparison, d_t preserves the uncertainty and partial agreement encoded in the full context distributions. Let $\rho(\mathbf{p}) = \max_k [\mathbf{p}]_k$ denote confidence of a context prediction. A memory read is rejected when

$$z_t = \mathbf{1}[M_{t-1} \text{ exists}]\, \mathbf{1}[\min\{\rho(\mathbf{p}_t), \rho(\mathbf{p}_M)\} \geq \kappa]\, \mathbf{1}[d_t \geq \tau], \tag{6}$$

where τ is selected on the validation set and κ optionally suppresses decisions based on low confidence context predictions. If either context distribution is unavailable, the predicted policy leaves the original temporal path unchanged. During inference, the policy relies exclusively on predicted context distributions.

The firewall decision directly determines routing:

$$(\ell_t, Q_t) = \begin{cases} F_\theta(x_t; M_{t-1}), & z_t = 0, \\ F_\theta(x_t; \varnothing), & z_t = 1. \end{cases} \tag{7}$$

When the memory is accepted, the model follows the original temporal path. When it is rejected, the memory is excluded from the current computation and therefore cannot influence either the output logits or the state candidate for that frame. Crucially, when a memory read is rejected, the image-only path produces a history-independent recovery candidate Q_t^A. If this candidate is subsequently admitted by the writer, it replaces the incompatible memory with a new state derived solely from the current scene. This recovery mechanism differs from write blocking, which simply retains the previous state and may allow an existing failure to persist. Thus, the firewall both prevents incompatible memory from influencing the current prediction and provides a candidate for recovery, while the writer independently determines whether that candidate should be committed to memory.

The firewall evaluates semantic compatibility rather than overall state quality. A high-quality memory originating from a different procedure may still exhibit a large d_t, whereas a visually degraded state from the correct procedure may receive a lower quality score while remaining semantically compatible. Keeping these two signals separate prevents the firewall from degenerating into an indiscriminate image-only fallback.

3.4 Trusted Memory Writer

The state produced by Eq. 7 may still be unreliable, even when its provenance is compatible. To address this problem, we introduce a decode-aware writer that evaluates the reliability of the routed state using information available after segmentation. For each query q, we construct the quality vector

$$\boldsymbol{\phi}_{tq} = [\, Q_{tq},\, \mathrm{Pool}(f_t; \mu_{tq}),\, \mathbf{c}_{tq}, \mathbf{s}_{tq}, \boldsymbol{\delta}_{tq}, \mathrm{sim}(Q_{tq}, M_{t-1,q}), a_{t-1}\,], \tag{8}$$

where f_t denotes the decoded image feature, μ_{tq} the predicted mask, $\mathbf{c}_{tq}$ summarizes class confidence, margin, and entropy, $\mathbf{s}_{tq}$ summarizes mask mean, variance, area, and boundary statistics, and $\boldsymbol{\delta}_{tq}$ captures class and mask disagreement with the image-only anchor. A lightweight MLP predicts a reliability score $\hat{u}_{tq}$ for each query, while a separate pooled head estimates the probability $\hat{b}_t$ that

the current frame is unreliable. We then aggregate the reliability scores for individual queries using active-query class confidence as weights:

$$s_t = \frac{\sum_q \omega_{tq} \hat{u}_{tq}}{\sum_q \omega_{tq}}, \tag{9}$$

where ω_{tq} denotes the predicted class confidence of foreground query q. If no foreground query is active, s_t is computed as the unweighted mean of the scores of each query.

During writer training, the segmentation backbone, decoder, and temporal pathway are frozen. We use matched soft Dice as the query-level reliability target u_{tq}. At the frame level, a prediction is labeled unreliable if its foreground Dice is below a predefined quality threshold or degrades sufficiently relative to the image-only anchor. The writer is trained with

$$\mathcal{L}_{\text{writer}} = \text{SmoothL1}(\hat{\mathbf{u}}_t, \mathbf{u}_t) + \lambda_b \, \text{BCE}(\hat{b}_t, b_t). \tag{10}$$

We fit a temperature T and bias β on validation and obtain the calibrated admission score

$$r_t = \sigma\left(\frac{\text{logit}(s_t)}{T} + \beta\right). \tag{11}$$

Given a validation-selected threshold γ, the sensory state is always refreshed with Q_t, whereas the trusted state is updated only when $r_t \geq \gamma$:

$$(M_t, \mathbf{p}_M, a_t) = \begin{cases} (Q_t, \mathbf{p}_t, 0), & r_t \geq \gamma, \\ (M_{t-1}, \mathbf{p}_M, a_{t-1} + 1), & r_t < \gamma. \end{cases} \tag{12}$$

As a result, the admission score r_t measures the reliability of a candidate for future memory and is distinct from the compatibility score d_t used by the firewall.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. ATLAS-120K [10] comprises more than 120,000 annotated frames, 30 semantic classes, and 14 surgical procedures. We follow the official split for training, validation, and testing. The test set contains 95 clips from 20 videos and covers 9 observed procedures. To evaluate our method under a different definition of semantic context, we use CholecSeg8k [18], which contains 8,080 annotated frames from 17 cholecystectomy videos and 13 segmentation classes. We align these frames with the 7 Cholec80 surgical phases [43], and split videos into 13/2/2 for training/validation/test. Its test split has 1,040 frames in 13 clips. Accordingly, "cross-context" refers to cross-procedure transfer on ATLAS and cross-phase transfer on CholecSeg8k. We report split-level Dice, referred to as Global Dice in our results. It is computed by first averaging frame-level class Dice within each class and then averaging across classes. Core-stress performance is defined as the unweighted mean of Global Dice across state-channel permutation, same-context history substitution, cross-context history substitution, and history dropout. Full-stress performance additionally includes forced reset, visual occlusion, and frame gap. The clean condition is excluded from both aggregates.

Controlled history interventions. For each intervention, we keep the current frame and its segmentation target fixed while modifying only the history state provided to the model. The core interventions include permutation of query order or state channels, history substitution using a state from another clip of the same context, history substitution using a state from a different context, and temporary history dropout. The full intervention set further includes forced reset, 15% visual occlusion, and a two-frame history gap.

Selection and uncertainty. All thresholds and adapter checkpoints are selected only on the validation set and fixed before testing. For all datasets and models, we use three pre-specified intervention seeds to assess sensitivity to randomness in the history interventions, and report the mean and standard deviation across seeds. These seeds affect only the intervention sampling and do not correspond to training. We further quantify uncertainty using paired cluster bootstrap with 10,000 replicates,

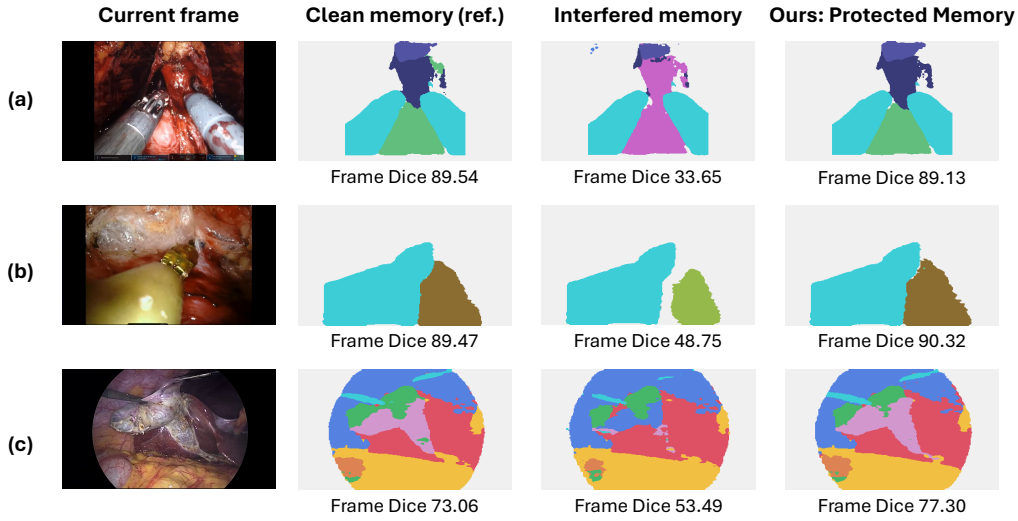

Figure 3: **Qualitative comparison under memory interference.** Each row shows the current frame, the reference prediction with clean memory, the prediction with interfered memory, and the corresponding result with our protected memory mechanism. Results are shown for (a) the original query memory model on ATLAS-120K, (b) TMANet on ATLAS-120K, and (c) ConvGRU on CholecSeg8k. Our method mitigates the degradation caused by corrupted temporal context and recovers predictions close to the clean memory reference.

Table 1: **Results across datasets and temporal models.** Values are Global Dice (%). Core and Full report gains under the corresponding stress conditions, and Cross denotes cross-procedure history on ATLAS and cross-phase history on CholecSeg8k.

		Clean			Stress gain		Cross-context history		
Dataset	Model	Base	Protected	Δ	ΔCore	ΔFull	Base	Protected	Δ
ATLAS	Query memory	31.53	31.66	+0.13	+1.76	+10.28	21.53	**28.59**	**+7.06**
	ConvGRU	28.52	28.52	< 0.01	+0.35	+0.20	26.79	**28.17**	**+1.38**
	TMANet	28.04	28.04	< 0.01	+0.49	+0.28	26.31	**28.15**	**+1.84**
CholecSeg8k	Query memory	44.95	44.95	< 0.01	+2.24	+1.40	41.58	**48.64**	**+7.06**
	ConvGRU	47.52	47.52	< 0.01	+1.85	+1.04	41.24	**48.65**	**+7.42**
	TMANet	48.54	48.54	< 0.01	+0.18	+0.10	47.92	**48.65**	**+0.73**

resampling test videos with replacement. For each replicate, we compute the paired change in mean frame-level Dice between PCMF and the corresponding baseline. The 2.5th and 97.5th percentiles are reported as the 95% confidence interval (CI).

4.2 Main Results

Table 1 compares the original temporal models with the same models equipped with PCMF. Across all six dataset–model pairs, the protected model consistently improves Global Dice under cross-context history, with gains ranging from 0.73 to 7.42. This improvement is observed under both procedure-level and phase-level context definitions. These consistent gains demonstrate the effectiveness of PCMF across diverse temporal memory mechanisms and semantic context definitions.

The gains are concentrated under history perturbations while the performance on clean mode remains essentially unchanged. Changes on the clean mode never exceed 0.13 and remain below 0.01 for other pairs, whereas every Core and Full stress aggregate improves. Moreover, the cross-phase gains on CholecSeg8k remain positive across all three intervention seeds, ranging from 7.365 to 7.467 for ConvGRU and from 0.7296 to 0.7306 for TMANet. These results indicate that the robustness gains are not obtained by unnecessarily suppressing useful temporal history under clean conditions, nor are they driven by a favorable intervention sample.

The convergence of protected scores across CholecSeg8k models reflects the effectiveness of PCMF in rejecting harmful cross-phase memory. Once incompatible memory is identified, they are consistently

Table 2: **Intervention analysis on the ATLAS test set.** Dice (%) of protected model combining PCMF, image-only recovery, and the Trusted Memory Writer.

Mode	Base	Protected	Δ	Trigger
Clean	31.53	31.66	+0.13	0.24%
Query permutation	31.00	31.20	+0.20	0.29%
Cross-proc. swap	21.53	**28.59**	**+7.06**	84.20%
History dropout	29.02	28.99	-0.03	0.41%
Forced reset	31.52	31.49	-0.03	0.29%
Visual occlusion	29.48	29.58	+0.10	4.67%
Frame gap	31.69	31.76	+0.07	0.43%

Table 3: **Sensitivity to the compatibility threshold on ATLAS validation.** All rows use temporal weight 0.4 and current-derived safe recovery. The bold threshold is the frozen operating point.

τ_r	Clean	Core	Cross Δ	Trigger	
				Clean	Cross
0.05	31.522	30.601	+6.314	3.8%	100.0%
0.10	31.680	30.788	+6.297	0.6%	99.2%
0.20	31.534	30.729	+6.074	0.2%	91.2%
0.30	31.912	30.819	+6.168	0.2%	89.2%
0.40	31.920	30.762	+6.136	0.1%	86.5%

Table 4: **Video-level bootstrap analysis on the ATLAS test set.** Results report paired changes in mean frame-level Dice (%) with 95% confidence intervals estimated from 10,000 video-level bootstrap resamples.

Model	Clean Δ [95% CI]	Cross-procedure Δ [95% CI]
Query memory	+0.0440 [-0.0530,+0.1540]	**+6.3930 [+4.6260,+8.5470]**
ConvGRU	-0.0003 [-0.0010, 0.0000]	**+0.6660 [+0.3530,+0.9910]**
TMANet	-0.0005 [-0.0016, 0.0000]	**+1.3200 [+0.8440,+1.8220]**

excluded from the read path, preventing them from contaminating the current prediction and allowing the models to rely on the image-only anchor instead. As a result, the protected models converge toward similar performance despite their substantially different unprotected scores. This behavior indicates that PCMF effectively prevents harmful temporal memory from influencing the current prediction across different memory mechanisms. In contrast, visual occlusion remains the most challenging intervention for the temporal models, indicating that the protection is specifically designed to address incompatible memory rather than general visual corruption.

4.3 Robustness Analysis

The firewall selectively rejects harmful history. The native ATLAS model provides the clearest view of the underlying mechanism, as Table 2. Cross-procedure history causes the largest degradation, reducing Global Dice to 21.53, while protection yields the largest improvement of +7.06. Its effects under history dropout, forced reset, visual occlusion, and frame gaps remain comparatively small. These results show that PCMF responds selectively to harmful cross-context memory rather than broadly suppressing temporal information.

The cross-context gain remains robust to video-level dependence. Treating frames as independent would underestimate uncertainty because neighboring surgical frames are strongly correlated. We therefore perform paired cluster bootstrap at the video level as Table 4. The 95% confidence interval for the cross-procedure gain remains strictly above zero for all three temporal models, whereas the corresponding clean interval includes or reaches zero. These results show that PCMF provides substantial robustness gains under incompatible memory.

4.4 Ablation Studies

Robustness is insensitive to the choice of compatibility criterion. Table 5 compares different rules for measuring the compatibility between the current scene and stored memory, while keeping the temporal weight and recovery strategy fixed. Both hard argmax and JS divergence substantially recover the performance loss under cross-procedure history, improving Dice by 6.10 and 6.17. The comparable performance of hard argmax and JS indicates that the robustness gain does not depend on a specific form of compatibility comparison. We adopt JS divergence because it preserves predictive uncertainty and reduces clean trigger rate from 4.3% to 0.2%, while maintaining a similar cross procedure trigger rate. Its main advantage is therefore more selective detection of incompatible memory rather than higher segmentation accuracy.

Table 5: **Compatibility criterion ablation on ATLAS.** All criterions use temporal weight 0.4 and safe recovery. GT context uses ground-truth procedure identity as reference. Trigger is the percentage of eligible memory reads rejected.

Criterion	Clean Dice	Cross Dice	Cross Δ	Clean Trigger	Cross Trigger
Base	31.58	23.92	+0.00	0.0%	0.0%
GT context	31.63	30.26	+6.34	0.0%	100.0%
Hard argmax	31.62	30.02	+6.10	4.3%	89.7%
JS, $\tau_r = 0.3$	31.91	30.09	+6.17	0.2%	89.2%

Table 6: **Effect of temporal reliance on PCMF.** Results on ATLAS use $\tau_r = 0.3$ and image-only recovery. Core is the mean over four history interventions. Worst is the minimum.

α	ΔClean	ΔCore	Worst-mode Dice		Cross Δ
			Monitor	Protected	
0.2	+0.15	+0.21	29.32	30.20	+0.88
0.3	+0.19	+0.66	27.37	30.15	+2.78
0.4	+0.33	+1.53	23.92	30.09	+6.17

The benefit of PCMF increases with temporal reliance. If PCMF improves robustness by rejecting incompatible temporal information, its benefit should become more pronounced as the model places greater weight on history. Table 6 evaluates this effect by varying the temporal fusion weight α while keeping the compatibility estimator and rejection threshold fixed. As α increases from 0.2 to 0.4, the base becomes increasingly vulnerable to harmful history, with decreasing from 29.32 to 23.92. In contrast, the protected model remains stable at approximately 30. Accordingly, the cross-procedure gain increases monotonically from 0.88 to 6.17, while the Core gain rises from 0.21 to 1.53. This consistent trend shows that PCMF becomes increasingly effective as the model relies more heavily on temporal memory.

PCMF is robust to the rejection threshold. Table 3 examines τ_r over an eightfold range, with the fixed temporal model and recovery strategy. The cross-procedure gain remains between 6.01 and 6.31, while Core Dice varies by only 0.22, indicating low sensitivity to the choice of threshold. Lower thresholds reject more incompatible history but also increase clean triggers, while higher thresholds reduce clean triggers but miss more cross-procedure mismatches. This analysis was conducted after threshold selection and is therefore used only to assess sensitivity. At $\tau_r = 0.30$, write blocking and image-only recovery achieve the same cross-procedure gain, while the latter improves clean Dice from 31.77 to 31.91. This result supports reconstructing the memory state from the current frame rather than retaining stale history.

5 Conclusion and Discussion

Conclusion. We identify provenance incompatibility as a distinct failure mode in stateful video segmentation, where plausible temporal memory can become harmful when its semantic context conflicts with the current scene. We propose PCMF to associate readable memory with predicted context provenance, verify compatibility before use, and reconstruct the state from the current frame when incompatible history is rejected. A Trusted Memory Writer further separates decoded sensory states from trusted memory retained for future reads. Across two surgical datasets and three temporal architectures, PCMF consistently improves robustness to cross-context history while preserving clean performance. Video-level bootstrap confirms that these gains persist across test videos. These results establish provenance verification as a general principle for reliable temporal inference.

Discussion. The primary contribution of PCMF is improved robustness to incompatible temporal memory while preserving performance on clean streams. Procedure and phase labels also provide only coarse provenance, which may be insufficient for gradual transitions, unseen contexts, or within-context changes in viewpoint and surgical stage. Image-only recovery limits the effect of incompatible history but cannot compensate for severe visual corruption, where reliable temporal evidence remains important. Future work should explore more precise provenance modeling, adaptive recovery, naturally occurring state failures, and the trade-off between robustness and deployment latency.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and code debugging. All technical content, experimental design, results, and interpretations were verified by the authors.

REFERENCES

[1] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[2] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 221–230, 2017.

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

[5] 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*, pp. 640–658. Springer, 2022.

[6] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5555–5564. IEEE, 2021.

[7] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in neural information processing systems*, 34:11781–11794, 2021.

[8] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161. IEEE, 2024.

[9] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pp. 343–352. Springer, 2020.

[10] Ronald LPD de Jong, Tim JM Jaspers, Raf AH Vervoort, Aron FHA Bakker, Yiping Li, Jip L Tolenaar, Jelle P Ruurda, Willem M Brinkman, Josien PW Pluim, Marcel Breeuwer, et al. Surgical anatomy recognition with context learning using foundation representations. *arXiv preprint arXiv:2606.22124*, 2026.

[11] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20167–20177. IEEE, 2023.

[12] 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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13614–13624. IEEE, 2025.

[13] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video cnns through representation warping. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4463–4472. IEEE, 2017.

[14] Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *International conference on medical image computing and computer-assisted intervention*, pp. 593–603. Springer, 2021.

[15] Maria Grammatikopoulou, Ricardo Sanchez-Matilla, Felix Bragman, David Owen, Lucy Culshaw, Karen Kerr, Danail Stoyanov, and Imanol Luengo. A spatio-temporal network for video semantic segmentation in surgical videos. *International Journal of Computer Assisted Radiology and Surgery*, 19(2):375–382, 2024.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

[17] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13434–13446. IEEE, 2023.

[18] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

[19] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8815–8824. IEEE, 2020.

[20] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8858–8867. IEEE, 2019.

[21] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan De Geus. Your vit is secretly an image segmentation model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25303–25313. IEEE, 2025.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

[23] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3168–3175, 2016.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[25] Jiangwei Lao, Weixiang Hong, Xin Guo, Yingying Zhang, Jian Wang, Jingdong Chen, and Wei Chu. Simultaneously short-and long-term temporal modeling for semi-supervised video semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14763–14772. IEEE, 2023.

[26] Sohyun Lee, Yeho Gwon, Lukas Hoyer, Konrad Schindler, Christos Sakaridis, and Suha Kwak. Robust promptable video object segmentation. *arXiv preprint arXiv:2605.12006*, 2026.

[27] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5997–6005. IEEE, 2018.

[28] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[29] Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931*, 2024.

[30] Haofeng Liu, Mingqi Gao, Xuxiao Luo, Ziyue Wang, Guanyi Qin, Junde Wu, and Yueming Jin. Resurgsam2: Referring segment anything in surgical video via credible long-term tracking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 435–445. Springer, 2025.

[31] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision*, pp. 352–368. Springer, 2020.

[32] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *European conference on computer vision*, pp. 661–679. Springer, 2020.

[33] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9): 691–696, 2017.

[34] Zhen-Liang Ni, Gui-Bin Bian, Zeng-Guang Hou, Xiao-Hu Zhou, Xiao-Liang Xie, and Zhen Li. Attention-guided lightweight network for real-time segmentation of robotic surgical instruments. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 9939–9945. IEEE, 2020.

[35] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6819–6828. IEEE, 2018.

[36] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9225–9234. IEEE, 2019.

[37] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[38] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

[39] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

[40] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *European Conference on Computer Vision*, pp. 629–645. Springer, 2020.

[41] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pp. 624–628. IEEE, 2018.

[42] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[43] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[44] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5272–5281. IEEE, 2019.

[45] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24255–24264. IEEE, 2025.

[46] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9473–9482. IEEE, 2019.

[47] Hao Wang, Weining Wang, and Jing Liu. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2254–2258. IEEE, 2021.

[48] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1296–1305. IEEE, 2021.

[49] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pp. 332–348. Springer, 2020.

[50] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in neural information processing systems*, 34:2491–2502, 2021.

[51] Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. Rmem: Restricted memory banks improve video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18602–18611. IEEE, 2024.

[52] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4141–4150. IEEE, 2017.

A Reproducibility and Evaluation

This appendix expands the experimental setting and reports results omitted from the main paper for space. All checkpoints, and recovery rules were selected on validation data and frozen before the corresponding formal test. The intervention seeds alter only the sampled history donor or permutation; they are not independent training seeds. Test-time ground-truth procedure or phase labels are never used by the predicted PCMF policy.

A.1 Dataset splits and evaluation units

ATLAS-120K. We use the official training, validation, and test split. The test split contains 27,289 annotated frames grouped into 95 clips from 20 videos and covers nine observed procedures. The native query-memory model and both temporal adapters share the same frozen image encoder and segmentation checkpoint. The formal test uses intervention seed 0; validation robustness is additionally checked with seeds 1 and 2 before locking.

CholecSeg8k. We align CholecSeg8k frames with the seven Cholec80 phase labels and split the 17 videos into 13/2/2 train/validation/test videos. The test split contains 1,040 frames in 13 clips from `video12` and `video27`. ConvGRU and TMANet formal tests use intervention seeds 0, 1, and 2. The native query-memory formal test uses the seed 0 .

Evaluation units. Global Dice averages frame-level Dice within each class and then across classes. Core and Full stress are unweighted means over the corresponding intervention sets, with Full additionally including forced reset, 15-frame visual occlusion, and a two-frame history gap. Cluster bootstrap reports paired changes in mean-frame Dice, which are distinct from Global Dice and reported separately.

A.2 Controlled history interventions

Every paired intervention keeps the current image, target mask, base checkpoint, and temporal-adapter checkpoint fixed. Only the supplied history or its availability is modified. Table 7 records the causal variable changed by each mode.

The same-context and cross-context interventions draw from disjoint donor pools defined before evaluation. Context provenance is stored with the readable memory and is updated atomically with that memory. Frames without a valid prior state are ineligible for a firewall decision and use the image-only anchor. Trigger rates therefore use eligible memory reads as the denominator.

A.3 Paired cluster bootstrap

We perform 10,000 paired bootstrap replicates. In each replicate, videos (20 clusters) or clips (95 clusters) are sampled with replacement, and both methods in a comparison share the identical resampled clusters. Confidence intervals are obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. The random seeds are fixed to 20270829 for the native query-memory experiment and 20270904 for the ATLAS ConvGRU and TMANet experiments. Bootstrap resampling is performed on saved paired frame records without additional inference.

B Temporal Model Implementations

B.1 Native query memory

The native ATLAS model propagates decoded mask queries. We use the frozen checkpoint selected before the firewall experiments and a fixed temporal fusion weight $\alpha = 0.4$ on ATLAS. The CholecSeg8k phase experiment uses the corresponding frozen segmentation checkpoint and its locked fast-memory path. Rejection removes the candidate history from the current computation; safe recovery then constructs a new candidate from the current image-only path.

Table 7: Controlled intervention suite. "Context" denotes procedure on ATLAS and phase on CholecSeg8k. Donor selection is deterministic conditional on the intervention seed.

Mode	Changed variable	Purpose
Clean	None	Measures ordinary temporal inference and the clean guardrail.
Query/state permutation	Query order for query memory, or hidden-state channels for recurrent adapters	Tests sensitivity to state identity or channel correspondence without changing the current frame.
Cross-context swap	History donor from a different procedure or phase	Isolates provenance-incompatible but otherwise plausible temporal evidence.
History dropout	Readable history removed for the current step	Measures behavior when temporal evidence is unavailable.
Forced reset	Temporal state reinitialized	Tests recovery from a stream reset.
Visual occlusion	Current observation occluded for 15 consecutive frames	Tests image degradation; this is not the primary target of PCMF.
Frame gap	Two intermediate history steps omitted	Tests temporal discontinuity without changing the current target.

B.2 Context prediction and computational scope

The current-frame provenance distribution is obtained from the image-only branch, before temporal fusion. Stored provenance is the distribution saved when the readable state was admitted. Thus a state cannot influence the evidence used to judge its own compatibility. The firewall itself introduces no trainable parameters; its additional work is the image-only provenance probe and, only on rejection, the already defined image-only recovery path.

C Complete Formal-Test Results

C.1 Per-mode results on ATLAS

Table 8 reports the complete paired Global Dice results across the three temporal mechanisms. Each comparison uses the same underlying model with and without firewall routing, including the shared procedure standard for the native query-memory model and the corresponding frozen adapters for ConvGRU and TMANet.

Table 8: Complete ATLAS formal-test results. Each cell is baseline/protected Global Dice (%). The parenthesized value is protected minus baseline in Dice points.

Model	Clean	Permute	Cross proc.	Dropout	Reset	Occlusion	Gap
Query memory	31.532/31.655 (+0.123)	31.000/31.201 (+0.201)	21.532/28.588 (+7.057)	29.018/28.985 (−0.032)	31.515/31.485 (−0.029)	29.482/29.577 (+0.095)	31.691/31.764 (+0.073)
ConvGRU	28.524/28.524 (< 0.001)	28.038/28.038 (< 0.001)	26.790/28.169 (+1.379)	28.465/28.465 (< 0.001)	28.528/28.528 (< 0.001)	26.786/26.790 (+0.004)	28.527/28.528 (+0.001)
TMANet	28.038/28.037 (< 0.001)	27.550/27.550 (< 0.001)	26.308/28.146 (+1.838)	28.444/28.444 (0.000)	28.051/28.051 (< 0.001)	26.350/26.374 (+0.024)	28.096/28.095 (−0.001)

Across all models, the largest improvement occurs under cross-procedure history. Despite substantially different memory formulations, the firewall consistently responds to provenance mismatch while preserving performance under clean and non-context interventions. These results demonstrate that compatibility-aware routing provides architecture-independent protection for temporal memory.

C.2 Per-mode results on CholecSeg8k

Table 9 reports the complete CholecSeg8k formal-test results. ConvGRU and TMANet results are averaged over three intervention seeds for seeded history substitutions, while other modes are seed invariant. The native query-memory model follows the fixed seed-0 setting.

Table 9: Complete CholecSeg8k formal-test results. Each cell is baseline/protected Global Dice (%), followed by the paired change in Dice points.

Model	Clean	Permute	Cross phase	Dropout	Reset	Occlusion	Gap
Query memory	44.945/44.945 (0.000)	37.325/37.325 (0.000)	41.583/48.643 (+7.060)	48.552/48.552 (0.000)	45.596/45.596 (0.000)	36.415/37.249 (+0.834)	45.259/45.259 (0.000)
ConvGRU	47.520/47.520 (0.000)	47.726/47.726 (0.000)	41.237/48.653 (+7.416)	48.653/48.653 (0.000)	47.562/47.562 (0.000)	38.658/38.496 (−0.162)	47.524/47.524 (0.000)
TMANet	48.536/48.536 (0.000)	48.508/48.508 (0.000)	47.923/48.653 (+0.730)	48.653/48.653 (0.000)	48.538/48.538 (0.000)	39.500/39.497 (−0.003)	48.545/48.545 (0.000)

Under cross-phase history, both protected models converge to the image-only anchor because all eligible incompatible reads are rejected. This convergence shows that predicted compatibility enables

Table 10: Paired mean-frame Dice change (%) with 95% cluster-bootstrap intervals on ATLAS. These are not Global Dice intervals.

Model	Cluster unit	Clean Δ [95% CI]	Cross-procedure Δ [95% CI]
Query memory	Video	$+0.044\ [-0.053, +0.154]$	$+6.393\ [+4.626, +8.547]$
	Clip	$+0.044\ [-0.047, +0.151]$	$+6.393\ [+4.571, +8.370]$
ConvGRU	Video	$-0.0003\ [-0.0010, 0.0000]$	$+0.666\ [+0.353, +0.991]$
	Clip	$-0.0003\ [-0.0010, 0.0000]$	$+0.666\ [+0.392, +0.953]$
TMANet	Video	$-0.0005\ [-0.0016, 0.0000]$	$+1.320\ [+0.844, +1.822]$
	Clip	$-0.0005\ [-0.0015, 0.0000]$	$+1.320\ [+0.921, +1.769]$

different temporal architectures to consistently route through the same recovery path without affecting their clean predictions.

D Uncertainty and Intervention-Seed Stability

D.1 Video- and clip-level bootstrap

Table 10 complements the video-level summary in the main paper with clip-level intervals. The cross-procedure interval is strictly positive for every ATLAS temporal model under both clustering units, whereas clean effects remain close to zero.

The native model achieves video-level core and full stress gains of $+1.607\ [+1.162, +2.139]$ and $+0.915\ [+0.659, +1.223]$ Dice points. The corresponding gains are $+0.120\ [+0.008, +0.228]$ and $+0.069\ [+0.005, +0.131]$ for ConvGRU, and $+0.325\ [+0.183, +0.482]$ and $+0.188\ [+0.106, +0.278]$ for TMANet. These consistent improvements confirm that the robustness gains are not artifacts of treating correlated frames as independent samples.

D.2 Sensitivity to intervention sampling

Table 11 summarizes the available multi-seed evaluations. Native ATLAS results are obtained after configuration fixing, while ATLAS adapters are evaluated on validation. All settings maintain positive cross-context gains across every evaluated intervention seed.

Table 11: Cross-context Global Dice gain across intervention seeds.

Dataset	Model	Seed 0	Seed 1	Seed 2	Mean $\pm$ std
ATLAS	Query memory	6.17	5.42	4.95	5.51 ± 0.50
ATLAS	ConvGRU	0.48	0.87	0.72	0.69 ± 0.16
ATLAS	TMANet	0.92	1.10	1.13	1.05 ± 0.09

The seed checks vary intervention sampling rather than model initialization. They therefore assess robustness to intervention realization, not training variance across independently initialized models.