LEARNING FROM RECOVERY: LOCALIZED EPISODIC TRAJECTORY MEMORY FOR CLOSED-LOOP MULTICLASS DECISIONS UNDER BINARY FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

We study multiclass decision making when an agent acts before receiving feedback and an error reveals only that the emitted action was wrong, not the hidden correct class. Our primary question is whether a persistent memory bank can turn such binary-feedback experience into better future decisions and more effective bounded retries. We introduce *Localized Episodic Trajectory Memory* (LETM), which stores context-conditioned winner-over-rejected relations from successful recovery trajectories such as $a_1^- \to a_2^- \to a_3^+$. The recurrent representation core is frozen; the trajectory-learning path uses no BP/BPTT, autograd-based update, or optimizer. Two evidence layers support the claim. First, mechanism controls show that disabling or scrambling persistent memory sharply degrades both first-decision and five-attempt performance, while random no-repeat retries recover only part of the closed-loop benefit. Second, in a frozen 10-seed closeout on UCI-HAR, N-MNIST, DVS-Gesture, and SHD, adding localized trajectory memory to the Full memory controller improves mean $\mathrm{First@1}$ by $+2.96$, $+2.16$, $+3.18$, and $+4.32$ percentage points. The First gains are paired-significant on UCI-HAR, N-MNIST, and SHD; DVS-Gesture instead shows significant gains in five-attempt recovery and conditional-recovery AUC. Mean attempts decrease on all four tasks, and Local exceeds Full at all 20 predeclared dataset-by-exposure checkpoints. A Global-Pairwise control and representation diagnostics further show that reusable memory depends on contextual localization and retrieval geometry, not merely on storing class-pair counts.

1 INTRODUCTION

Interactive systems often receive evaluative feedback rather than full supervision. A robot may know that an attempted action failed; a recommender may learn that a chosen item was rejected; an online classifier may be told only whether its prediction was correct. Multiclass bandit learning formalizes a core version of this setting: the learner predicts a class and observes whether the prediction was correct without necessarily observing the true label (Kakade et al., 2008; Wang et al., 2010; Ngo et al., 2013; 2014; Chen et al., 2014; Zhang et al., 2019; Beygelzimer et al., 2019; van der Hoeven et al., 2021). Recent work also studies binary correctness feedback for test-time model adaptation (Lee et al., 2025).

Most such formulations organize learning across rounds. Here we focus on a complementary axis: *what happens inside one failed decision*. Suppose the learner can keep the same instance active after an error, produce another action, and continue until success or a small action budget is exhausted. A trajectory

$$a_1^- \to a_2^- \to \cdots \to a_k^+ \tag{1}$$

contains more structure than a scalar loss. Negative feedback certifies each a_i^- as a rejected alternative, while positive feedback on a_k^+ certifies that the learner's own emitted action is correct for that instance. The recovery therefore induces comparisons between the eventual winner and actions already known to fail.

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.

The central question is whether persistent memory can make these experiences useful beyond the current attempt. We separate two levels of evidence. At the *memory-bank* level, we ask whether persistent structured memory improves first decisions and feedback-assisted retries compared with disabling, scrambling, or randomizing memory. At the *trajectory-memory* level, we ask whether the particular winner-over-rejected relation created by successful recovery adds value on top of an already memory-equipped controller.

If Eq. (1) is discarded once the current instance is solved, recovery helps only that instance. If it is stored together with the problem representation, a later similar instance can retrieve the relation *before its first external action*. We introduce *Localized Episodic Trajectory Memory* (LETM), which stores

$$(z,\ a^+ \succ a_i^-) \tag{2}$$

for a frozen representation z after a successful recovery. At a future query, nearby trajectory relations provide a context-conditioned preference correction. We compare this localized memory against both the existing Full memory controller and a Global-Pairwise control that retains winner–loser statistics while discarding the semantic/context key.

Our protocol separates first-decision quality from closed-loop recovery:

$$\mathrm{First@1} = \Pr(a_1 = y), \qquad \mathrm{Solve@}K = \Pr(\exists t \le K : a_t = y). \tag{3}$$

$\mathrm{Solve@5}$ is *not* ordinary Top-5 accuracy: the learner receives a new binary outcome after each emitted action. This distinction lets us test both of the paper's practical claims: whether memory improves the first action, and whether it helps the system use repeated attempts more effectively when the first action is wrong.

The evidence is deliberately layered. In five-seed mechanism diagnostics, removing persistent memory causes large drops in First@1 and/or $\mathrm{Solve@5}$ on UCI-HAR, N-MNIST, and DVS-Gesture; random no-repeat retries and shuffled memory action associations do not recover the structured controller's performance. In the stricter frozen 10-seed closeout, localized trajectory memory then adds a further First@1 improvement over the Full memory controller on all four tasks: $+2.96$, $+2.16$, $+3.18$, and $+4.32$ pp on UCI-HAR, N-MNIST, DVS-Gesture, and SHD. For UCI-HAR, N-MNIST, and SHD, the paired First effect is positive in all ten seeds ($p = 0.001953$ under the exact sign-flip test). On DVS-Gesture, First is less stable, but $\mathrm{Solve@5}$ improves by 3.33 pp ($p = 0.0195$) and conditional-recovery AUC by 4.79 pp with all ten seeds positive. Mean attempts fall on all four tasks, and Local exceeds Full at all 20 combinations of four datasets and five exposure checkpoints ($0.25, 0.5, 1, 2, 4$ passes).

The experiments also expose a second result: episodic memory quality depends on the geometry of its key space. On DVS-Gesture, richer temporal representations greatly improve local retrieval and closed-loop behavior. A temporal-difference representation provides a counterexample: class-relevant information remains detectable by a centroid probe, yet nearest-neighbor purity and memory performance collapse. Thus "information is present" and "experience is retrievable where it is useful" are distinct properties.

Contributions.

- We formulate a same-instance binary-feedback protocol that evaluates memory on two distinct outcomes: future first-action accuracy and feedback-assisted repeated-attempt success.
- We provide direct mechanism evidence that a persistent structured memory bank materially contributes to both outcomes: removing or scrambling memory degrades performance, and random no-repeat retries explain only part of the closed-loop gain.
- We introduce LETM, a localized trajectory-memory extension that stores externally verified winner-over-rejected relations. A frozen 10-seed closeout shows additional First@1 gains over the already memory-equipped Full controller, with reduced mean attempts on all four datasets.
- Global-Pairwise, exposure, causal-control, capacity, source/protocol, and representation diagnostics isolate why memory helps: contextual localization matters, and local addressing geometry can fail even when class information remains detectable.

2 Related Work

Multiclass partial and bandit feedback. Banditron (Kakade et al., 2008) and subsequent partial-feedback algorithms (Wang et al., 2010) learn online when the oracle reveals only whether the predicted class is correct. UCWL explicitly targets multiclass binary right/wrong feedback and human–robot interaction (Ngo et al., 2013; 2014). Later work develops boosting and efficient linear algorithms for multiclass bandit prediction (Chen et al., 2014; Zhang et al., 2019; Beygelzimer et al., 2019), while feedback-graph formulations generalize the information revealed after a prediction (van der Hoeven et al., 2021). Binary correctness feedback is therefore not itself our novelty; our focus is the within-instance recovery trajectory and the persistent, context-localized relation it leaves behind.

Binary-feedback and backpropagation-free adaptation. BiTTA (Lee et al., 2025) adapts pre-trained models under domain shift using sparse binary correctness feedback. T3A (Iwasawa & Matsuo, 2021) performs backpropagation-free test-time classifier adjustment using pseudo-prototypes, AdaNPC (Zhang et al., 2023) uses non-parametric feature–label memory, and FOA (Niu et al., 2024) performs test-time adaptation using only forward passes and derivative-free optimization. These works motivate memory- or forward-based adaptation, but they do not make the recovery trajectory in Eq. (1) the persistent learning object.

Episodic control. Model-Free Episodic Control (Blundell et al., 2016) and Neural Episodic Control (Pritzel et al., 2017) show that rapidly written episodic memories can support efficient sequential decisions. Our feedback is weaker and our stored object is more specific: a context-localized comparison between an action externally rejected on the current problem and an action later confirmed correct. Banditron/UCWL update across examples after a prediction, whereas MFEC/NEC use reinforcement-learning return/value targets; their published numbers are therefore not protocol-matched numerical baselines for our same-instance binary-feedback loop. We instead use identical-encoder, identical-split mechanism controls and report differently supervised external systems as contextual references.

3 Problem Setting

Let $\mathcal{A} = \{1, \ldots, C\}$ be the action/classes. Each instance x has an environment-owned correct class y that is not passed to the predictor. A frozen encoder produces $z = \phi(x)$. At external attempt t,

$$a_t = F(z, M, A_{<t}), \tag{4}$$

where M is persistent memory and $A_{<t}$ records actions already attempted on the current instance. Only after the action is emitted does the environment reveal

$$r_t = \begin{cases} +1, & a_t = y, \\ -1, & a_t \neq y. \end{cases} \tag{5}$$

After $r_t = -1$, y remains hidden. The episode terminates on the first positive outcome or after $K = 5$ external actions. Across learning instances $x^{(1)}, x^{(2)}, \ldots$, the predictor reads persistent state M_n before acting on instance n; feedback may change within-instance rejection/failure state, and a verified successful recovery may update persistent memory to M_{n+1}. During official test, persistent state is frozen across instances ($M_{n+1} = M_n$), while within-instance feedback/retry remains active. Because the four tasks have $C = 6, 10, 11, 20$ classes, the fixed $K = 5$ budget represents different fractions of the action space; we therefore do not interpret cross-dataset $\mathrm{Solve@5}$ values as equally difficult classification scores, and report $\mathrm{First@1}$ plus no-repeat controls separately. For initially wrong cases we also report conditional recovery

$$R_K = \frac{\mathrm{Solve@}K - \mathrm{First@1}}{1 - \mathrm{First@1}}, \tag{6}$$

and summarize $R_2, \ldots, R_5$ by their mean ("recovery AUC"; not ROC AUC).

4 Method

4.1 Frozen recurrent representation and decision dynamics

All Full/Global/Local comparisons share the same recurrent representation core (nominal hidden size 2048) and the same base decision controller. For sequential input $x_{1:T}$ the encoder produces H_T

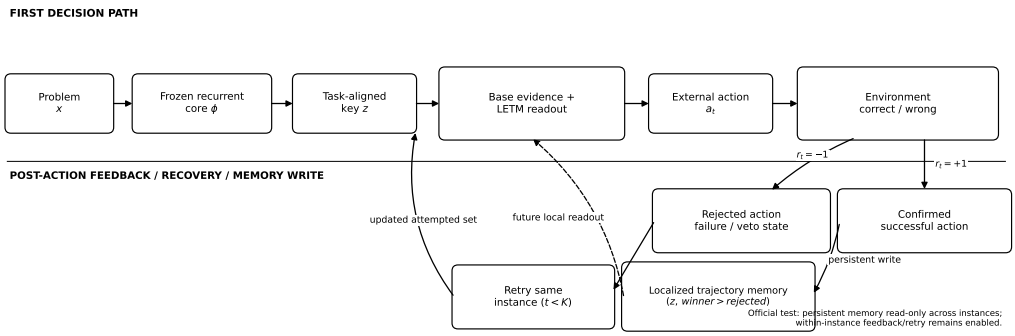

Figure 1: **Closed-loop learning with LETM.** A frozen recurrent core produces the addressing representation. Negative feedback records rejected alternatives and permits same-instance retry. Only after positive feedback is the successful action paired with previously rejected actions and written to localized trajectory memory. During official test, persistent memory is read-only across test instances.

and is frozen throughout the memory-learning experiments. Freezing ϕ is an experimental control: it isolates what persistent binary-feedback memory contributes without conflating the effect with representation retraining.

The controller forms a class-evidence vector $d \in \mathbb{R}^C$. Before the LETM correction, the audited base drive is

$$s_c^{base} = 0.42 s_c^{core} + 3.20 s_c^{positive} - 4.00 s_c^{veto} + 2.40 s_c^{stable} + 0.90\, \mathbf{1}_{retry} s_c^{candidate} - 20\, \mathbf{1}[c \in A_{rej}], \tag{7}$$

where $s_c^{core} = z^\top w_c$ is the frozen linear readout; $s^{positive}$ and s^{veto} are the persistent positive/failure-memory readouts; and s^{stable} and $s^{candidate}$ are the stable-theory and retry-candidate scores. The candidate term is active only after a failed external action, while the attempted-action penalty implements exclusion. For Full, $d = s^{base}$; for Local, LETM adds the trajectory term defined below. The effective drive is initialized by $q_0 = \tanh(d)$ and then uses the audited bounded action-space settling recurrence

$$q_{\tau+1} = \tanh\big(0.70 q_\tau + 0.30 d + \gamma(q_\tau W^{act})\big)\,, \tag{8}$$

where $q_\tau, d \in \mathbb{R}^C$, $W^{act} \in \mathbb{R}^{C \times C}$ is the fixed action-interaction matrix, and $\gamma = 0.10$ is the fixed recurrent-action gain. Settling runs for at most 48 internal cycles (with a minimum of four before early stopping); the emitted action is the highest-scoring non-excluded class, and rejected actions are excluded on retries. These settling cycles are distinct from recurrence through the 2048-unit encoder. LETM changes only the evidence drive through an additional recovery-derived readout.

4.2 Failure becomes a rejected alternative

If $r_t = -1$, then $a_t \neq y$ is a valid constraint. The action is added to the current rejected/attempted set. Prediction is side-effect free; persistent writes occur only after the outcome boundary. A failed action may update failure memory and current problem state but does not directly rewrite the stable theory. These ordering constraints were checked by static and runtime protocol audits.

4.3 Localized Episodic Trajectory Memory

Suppose an episode ends as $a_1^- \to a_2^- \to a_3^+$. After positive feedback on a_3, the system stores

$$(z, a_3 \succ a_1), \qquad (z, a_3 \succ a_2). \tag{9}$$

The write API receives only (z, successful action, rejected-action mask); it does not receive a target/label argument. Each winner has a bounded bank of contextual slots. Writing first L_2-normalizes z and compares it with the normalized slot centers by cosine similarity. A sufficiently similar winner-specific slot is merged; otherwise a free slot is opened, and if the bank is full the nearest slot is force-merged. The frozen closeout uses 32 slots per winner for UCI-HAR, N-MNIST, and DVS-Gesture and 16 for SHD; the 8/16/32/64 sweep is audit-only.

Algorithm 1 Learning from a recovery trajectory

Require: x, frozen encoder ϕ, memory M, max attempts K

$z \leftarrow \phi(x)$; $A_{\text{rej}} \leftarrow \varnothing$
for $t = 1, \ldots, K$ **do**
  $s \leftarrow s^{base}(z, M, A_{\text{rej}}) + \lambda s^{traj}(z, M)$
  $a_t \leftarrow \text{DecisionSettle}(s, A_{\text{rej}})$
  emit a_t; receive $r_t \in \{-1, +1\}$
  if $r_t = -1$ **then**
    $A_{\text{rej}} \leftarrow A_{\text{rej}} \cup \{a_t\}$; update failure state
  else
    for all $a^- \in A_{\text{rej}}$ **do**
      store $(z, a_t \succ a^-)$ if writes are enabled
    end for
    return success at attempt t
  end if
end for
return unsolved within K

For a normalized query z', let μ_{ik} be the normalized center of slot k for winner i, and let n_{ikj} be its stored count for relation $i \succ j$. The compatibility implementation uses

$$K_{ik}(z') = \exp\left[-\tfrac{1}{2}\left(\frac{[1 - z'^{\top}\mu_{ik}]_+}{0.095}\right)^2\right], \tag{10}$$

$$P_{ij}(z') = \max_{k \in \mathcal{V}_i} K_{ik}(z')\left(1 - e^{-n_{ikj}/2}\right), \tag{11}$$

with $P_{ii} = 0$ and $\mathcal{V}_i$ the valid slots for winner i. Pair evidence is reduced to a class-level preference by normalized outgoing-minus-incoming support,

$$s_c^{traj}(z') = \tanh\left(2.5\left[\frac{\sum_j P_{cj}(z')}{C-1} - \frac{\sum_i P_{ic}(z')}{C-1}\right]\right). \tag{12}$$

The frozen trajectory gain is $\lambda = 0.90$, fixed before the 10-seed closeout and never selected on the official test:

$$s_c^{\text{LETM}}(z') = s_c^{base}(z') + 0.90\, s_c^{traj}(z'). \tag{13}$$

The exact earliest historical trajectory source is not claimed to be byte-identical; these equations describe the audited compatibility implementation used for the reported closeout. The Global-Pairwise control keeps the same pair-to-class reduction but removes z and the contextual slots, leaving unconditional winner–loser counts.

4.4 Task-aligned addressing keys and evaluation boundary

The learning protocol is fixed, but a memory can only retrieve experience through its key geometry. The frozen closeout uses: (i) UCI-HAR and N-MNIST, normalized unprojected $H_T \in \mathbb{R}^{2048}$; (ii) DVS-Gesture, the predeclared 128-D projected semantic key; and (iii) SHD, four separately encoded temporal prefixes $[H_{25\%}, H_{50\%}, H_{75\%}, H_{100\%}] \in \mathbb{R}^{8192}$. Representation choices are frozen before official-test evaluation. During learning, evaluated outcomes update persistent evidence and trajectory memory. The official test is touched once after pass 4; persistent state is read-only across official-test examples. Within-instance retry remains enabled because it defines $\text{Solve@}K$.

5 Experimental Protocol

We evaluate UCI-HAR (Anguita et al., 2013), N-MNIST (Orchard et al., 2015), DVS-Gesture (Amir et al., 2017), and SHD (Cramer et al., 2022). The frozen adapters use 6/10/11/20 classes, 32/34/40/40 time steps, and learn/audit/test sizes 5200/900/2947, 15000/2000/10000, 912/165/264, and 6000/1200/2264, respectively. All preprocessing is label-free at inference; UCI-HAR channel statistics and SHD global-duration statistics are training-derived. The main closeout compares **Full**,

Table 1: Five-seed mechanism diagnostic: First@1 / Solve@5 (%). These are causal diagnostics, not the frozen 10-seed headline comparison. SHD is omitted here because the earlier semantic key was later shown to be a severe bottleneck.

System	UCI-HAR	N-MNIST	DVS-Gesture
No persistent memory	19.82 / 84.77	8.65 / 50.59	10.61 / 47.58
Full structured memory	**77.43 / 98.64**	**85.03 / 98.99**	**39.32 / 86.21**
Random no-repeat retries	77.43 / 94.71	85.03 / 91.63	39.32 / 63.03
Memory–action shuffle	9.92 / 88.33	19.10 / 58.74	10.45 / 52.88

Global-Pairwise, and **Local-Pairwise (LETM)** under identical splits and a five-action budget. It uses ten seeds $\{7, 17, 27, 37, 47, 57, 67, 77, 87, 97\}$ and four cumulative passes through the learn split. The audit split is additionally read at $0.25, 0.5, 1, 2, 4$ passes; capacity sweeps are audit-only. All closeout comparisons are paired by seed with bootstrap 95% confidence intervals and exact two-sided sign-flip tests. A reconstruction audit verifies frozen source hashes, absence of gradient API calls in the reconstructed trajectory extensions, label-free prediction/write APIs, and action-before-feedback ordering; a separate benchmark audit verifies split isolation and preprocessing boundaries.

Evidence hierarchy and comparison logic. The frozen 10-seed *confirmatory* comparison is Full versus Global-Pairwise versus Local-Pairwise under identical splits, seeds, representations, and exposure. Here Full asks what the existing memory controller can do, Global asks whether unlocalized winner–loser counts are enough, and Local asks whether context-localized recovery relations add further value. Separate five-seed *mechanism diagnostics* disable persistent memory, remove attempted-action exclusion, replace retries by random no-repeat elimination, shuffle memory-to-action associations, or use random lookup. These diagnostics test causal alternatives and are not substituted for the frozen closeout. Appendix D reports frozen-representation BP/MLP, local-forward-style, reward-modulated Hebbian, and fully supervised end-to-end references; because their supervision/update regimes differ, they are contextual rather than protocol-matched baselines.

6 Results

6.1 Persistent structured memory improves decisions and retries

Before asking whether LETM improves an already memory-equipped controller, we first test whether structured persistent memory matters at all. Table 1 collects representative five-seed mechanism diagnostics on the three tasks whose keys were already usable in that evidence chain. Disabling persistent memory causes large losses in First@1 and/or Solve@5. Keeping the same first action but replacing structured retry with random no-repeat elimination recovers only part of Solve@5, while shuffling memory-to-action associations damages both first decisions and recovery. Thus the closed-loop gains cannot be reduced to "try a different class after failure."

This first result establishes the broad Paper-1 claim: a structured memory bank can improve both immediate decision quality and repeated-attempt success under binary feedback. The stricter question is then whether *recovery trajectories themselves* contribute additional reusable information beyond the existing Full memory system.

6.2 Localized recovery memory adds further first-decision gains

Local-Pairwise improves First@1 over Full on all four tasks. UCI-HAR, N-MNIST, and SHD are positive in all ten paired seeds, giving the minimum possible two-sided sign-flip $p = 0.001953$ for ten nonzero paired effects. DVS-Gesture is positive in 9/10 seeds but its First 95% CI crosses zero; its Solve@5 gain is $+3.33$ pp ($p = 0.0195$) and recovery-AUC gain is $+4.79$ pp ($p = 0.001953$). Appendix A reports the paired effect sizes and bootstrap 95% CIs, making the one noisy confirmatory result explicit rather than hiding it.

Table 2: Frozen 10-seed official closeout. Recovery AUC is mean conditional recovery over attempts 2–5. Solve@5 is feedback-assisted, not standard Top-5 accuracy.

Dataset	System	First@1	Solve@5	Mean attempts	Recovery AUC
UCI-HAR	Full	79.11	98.66	1.364	79.77
	Global	78.38	98.95	1.360	82.07
	Local (LETM)	**82.07**	98.70	**1.305**	80.63
N-MNIST	Full	90.54	99.40	1.160	81.07
	Global	90.56	99.42	1.162	80.56
	Local (LETM)	**92.70**	**99.70**	**1.114**	**84.78**
DVS-Gesture	Full	41.36	84.58	2.448	56.67
	Global	42.69	86.97	2.377	59.32
	Local (LETM)	**44.55**	**87.92**	**2.293**	**61.46**
SHD	Full	40.00	74.92	2.676	44.69
	Global	37.12	73.05	2.776	43.71
	Local (LETM)	**44.32**	**76.64**	**2.551**	44.90

6.3 Memory improves repeated attempts; trajectory memory shifts success earlier

The diagnostic results in Table 1 show that the broader structured memory system can strongly improve repeated-attempt success. Once that memory system is already present, the *incremental* gain from LETM is concentrated earlier in the trajectory. On UCI-HAR, Local-minus-Full changes from $+2.96$ pp at attempt 1 to $+0.04$ pp by attempt 5; N-MNIST changes from $+2.16$ to $+0.31$ pp, and SHD from $+4.32$ to $+1.72$ pp. DVS-Gesture is the exception: under its frozen semantic128 key, Local retains a larger $+3.33$ pp Solve@5 gain. Mean attempts nevertheless fall by approximately 4.31%, 3.96%, 6.33%, and 4.68% on UCI-HAR, N-MNIST, DVS-Gesture, and SHD. The interpretation is therefore two-level: persistent structured memory supports recovery in general, while localized trajectory memory mainly moves success toward the first action and reduces the number of external trials.

Localization and few-pass behavior. Local-Pairwise has higher First@1 than Global-Pairwise on all four datasets. The Local-minus-Global First effect is significant on UCI-HAR, N-MNIST, and SHD; DVS again has a higher mean with a wider seed distribution. Across $0.25, 0.5, 1, 2, 4$ passes, Local First exceeds Full at all 20 dataset-by-exposure checkpoints.

Recovery is not no-repeat alone. Removing attempted-action exclusion sharply reduces Solve@5 on UCI-HAR, N-MNIST, and DVS-Gesture, while random no-repeat elimination also underperforms the structured controller. On DVS-Gesture, shuffling memory-to-action associations or using random action lookup drops Solve@5 to roughly 0.53, compared with 0.86 for the structured Full controller in the five-seed evidence chain. Additional controls appear in Appendix C.

7 Addressing Geometry and Residual Failure

The DVS experiments expose an important limitation of episodic learning. The frozen closeout used a 128-D projected semantic key; a later five-seed diagnostic ladder evaluates label-free alternatives while keeping the closed-loop machinery fixed. Moving to the unprojected final 2048-D state improves nearest-neighbor retrieval and Local-Pairwise First, and four temporal prefixes improve further. The strongest diagnostic key, a fixed 512-D projection of raw temporal statistics, reaches about 67.27% First and 97.94% Solve@5.

The key counterexample is `prefix4delta2048`. Its centroid probe still detects substantial class-relevant structure (about 49.9% top-1), yet 1-NN accuracy falls to about 29.3% and the closed-loop system collapses to about 9.2% First. This pattern is consistent with a representation that preserves some global class separation while distorting local neighborhoods: temporal differencing can amplify

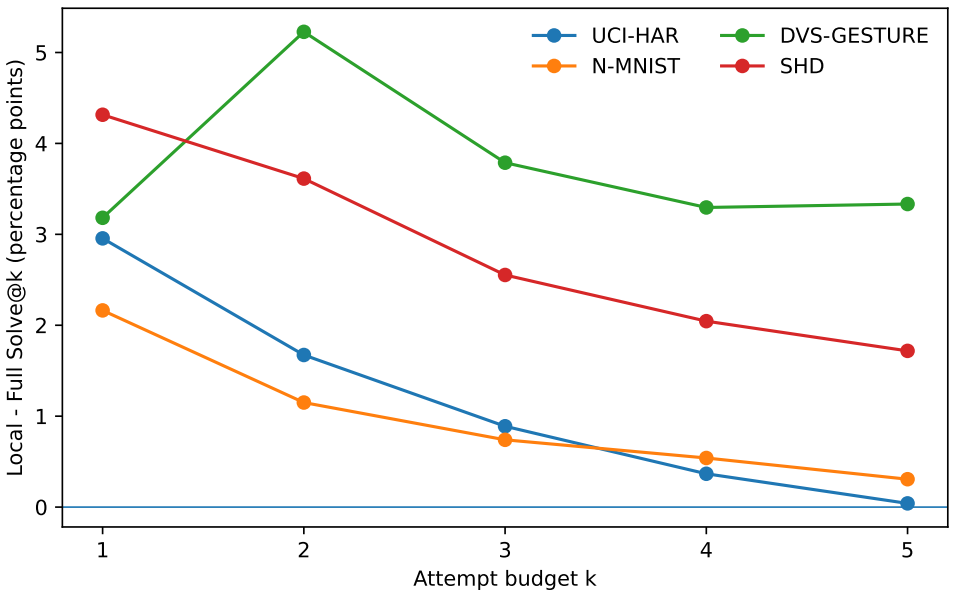

Figure 2: Local-minus-Full cumulative Solve@k difference. The benefit is predominantly an early-decision effect on UCI-HAR, N-MNIST, and SHD; DVS retains a larger recovery benefit under its frozen semantic128 key.

instance-specific variation, so globally detectable information need not remain locally transferable. We treat this as a mechanism hypothesis rather than a causal identification. Thus detectable class information does not imply retrieval-friendly episodic geometry.

A later five-view DVS diagnostic further separates candidate generation from candidate selection. The union of the views' Top-1, Top-2, and Top-3 candidates contains the correct class in 93.94%, 97.94%, and 99.52% of audit examples, while the best learned arbitration reaches 73.21% First. The residual DVS bottleneck is therefore dominated by discrimination among plausible candidates rather than candidate availability.

8 Discussion and Limitations

The results support three bounded conclusions. First, persistent structured memory is not a cosmetic component: mechanism controls show large losses when memory is disabled or its action associations are scrambled, and random no-repeat retries do not reproduce the same closed-loop performance. Second, recovery itself supplies an additional reusable memory object: a wrong action establishes a negative fact, a later positive action establishes a winner, and localizing that verified relation improves future first decisions on top of the Full memory controller. Third, memory and representation cannot be analyzed independently. An episodic learner is an addressing system; it succeeds only when the key space places future queries near experiences whose relations should transfer.

This is not a conventional SOTA-classification claim. End-to-end supervised BP/BPTT references can outperform LETM First on several tasks and use a stronger supervision/update regime. Our claim is instead about the weak-feedback protocol: persistent first-decision improvements can be obtained from the learner's own externally evaluated recovery trajectory while leaving the central recurrent core frozen.

The strongest confirmatory evidence is the frozen 10-seed closeout. Higher post-closeout audit-selected results (94.24% UCI-HAR, 93.46% N-MNIST, 73.21% DVS-Gesture, 62.37% SHD) are reported only in Appendix F and require independent confirmation if used as headline performance. DVS First remains seed-sensitive under the frozen semantic128 key; SHD remains representation-limited relative to supervised references. The evaluated tasks are temporal/sensor or neuromorphic benchmarks, so transfer to standard static-image domains remains open. The current work keeps

episodic memory external and bounded by fixed per-winner slots; long-horizon retrieval latency, alternative merge/forgetting policies, and memory-to-core consolidation are future questions. No physical-robot deployment is reported, where failed retries would also carry time, energy, safety, or user-experience costs. Finally, the earliest historical trajectory-preference source was not recovered byte-for-byte; the compatibility reconstruction preserves confirmed API/order invariants and was source-audited, while the recovered frozen core/evidence sources are hash-verified.

9 Conclusion

This paper tests a deliberately focused claim: persistent memory can make a binary-feedback closed-loop learner more accurate on its first action and more effective when repeated attempts are needed. Mechanism controls support the importance of structured memory itself, and LETM shows that successful recovery trajectories contain an additional reusable relation that improves future First@1 on top of an already memory-equipped controller. Across four benchmarks, localized recovery memory also lowers mean attempts, while representation diagnostics show that the usefulness of stored experience depends on retrieval geometry rather than merely on globally detectable class information. The result is a clean Paper-1 foundation: remember verified experience first; core consolidation can be studied separately.

Reproducibility Statement

The main paper specifies the interaction protocol, attempt budget, frozen representations, ten closeout seeds, exposure schedule, and paired statistical tests. Appendix H provides benchmark preprocessing, capacity settings, source/protocol audit outcomes, and the distinction between frozen official results and later audit-selected experiments. The experimental code and result tables are maintained as versioned artifacts; an anonymized code package can be supplied as supplementary material. The author is responsible for matching any released package to the frozen protocol described here.

AI Use Statement

Generative AI tools were used to help refine research hypotheses and experimental methodology, implement and debug portions of experimental code, interpret intermediate results, identify and summarize related literature, create or edit scientific figures, and draft and edit manuscript text. The experiments themselves were executed in the author's research environment, and AI-assisted claims were checked against stored result files, source/protocol audits, and cited sources. The author reviewed the final manuscript and takes responsibility for all claims, code, figures, citations, and other AI-assisted artifacts included in the submission.

Ethics Statement

This work introduces no new human-subject data collection and evaluates only established public benchmark datasets. It does not release personal data or deploy an autonomous system in a physical environment. The main foreseeable risk is over-interpreting feedback-assisted $\mathrm{Solve@}K$ as ordinary classification accuracy; the paper explicitly separates these metrics and reports supervision differences for baseline comparisons.

References

Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7243–7252, 2017.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st*

European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2013.

Alina Beygelzimer, David Pal, Balazs Szorenyi, Devanathan Thiruvenkatachari, Chen-Yu Wei, and Chicheng Zhang. Bandit multiclass linear classification: Efficient algorithms for the separable case. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 624–633, 2019.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z. Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. *arXiv preprint arXiv:1606.04460*, 2016.

Shang-Tse Chen, Hsuan-Tien Lin, and Chi-Jen Lu. Boosting with online binary learners for the multiclass bandit problem. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 342–350, 2014.

Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2022. doi: 10.1109/TNNLS.2020.3044364. Early access 2020.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 440–447, 2008. doi: 10.1145/1390156.1390212.

Taeckyung Lee, Sorn Chottananurak, Junsu Kim, Jinwoo Shin, Taesik Gong, and Sung-Ju Lee. Test-time adaptation with binary feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33005–33024, 2025.

Hung Ngo, Matthew Luciw, Ngo Anh Vien, and Jürgen Schmidhuber. Upper confidence weighted learning for efficient exploration in multiclass prediction with binary feedback. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 2488–2494, 2013.

Hung Quoc Ngo, Matthew D. Luciw, Jawad Nagi, Alexander Förster, Jürgen Schmidhuber, and V. Ngo. Efficient interactive multiclass learning from binary feedback. *ACM Transactions on Interactive Intelligent Systems*, 4(3):1–25, 2014. doi: 10.1145/2629631.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38298–38315, 2024.

Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015. doi: 10.3389/fnins.2015.00437.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2827–2836, 2017.

Dirk van der Hoeven, Federico Fusco, and Nicolò Cesa-Bianchi. Beyond bandit feedback in online multiclass classification. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shijun Wang, Rong Jin, and Hamed Valizadegan. A potential-based framework for online multi-class learning with partial feedback. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 900–907, 2010.

Daniel T. Zhang, Young Hun Jung, and Ambuj Tewari. Online multiclass boosting with bandit feedback. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1148–1156, 2019.

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. AdaNPC: Exploring non-parametric classifier for test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41647–41676, 2023.

A Evidence hierarchy and full paired statistics

The paper separates two result layers. The **confirmatory layer** is the frozen 10-seed closeout comparing Full, Global-Pairwise, and Local-Pairwise on all four datasets; official test is evaluated once after four passes and persistent memory is read-only across test examples. The **exploratory layer** contains later five-seed audit searches over representations, candidate restrictions, successful-episode posteriors, and DVS multi-view arbitration.

Table 3: Local-Pairwise minus Full, frozen 10-seed closeout.

Dataset	Metric	Mean Δ (pp)	Bootstrap 95% CI (pp)	Exact p
UCI-HAR	First@1	+2.96	[2.44,3.47]	0.001953
	Solve@5	+0.04	[-0.11,0.18]	0.640625
	Recovery AUC	+0.86	[-0.46,2.08]	0.240234
N-MNIST	First@1	+2.16	[1.89,2.46]	0.001953
	Solve@5	+0.31	[0.23,0.39]	0.001953
	Recovery AUC	+3.71	[2.69,4.82]	0.001953
DVS-Gesture	First@1	+3.18	[-0.11,6.06]	0.091797
	Solve@5	+3.33	[1.21,5.38]	0.019531
	Recovery AUC	+4.79	[2.96,6.63]	0.001953
SHD	First@1	+4.32	[2.87,5.77]	0.001953
	Solve@5	+1.72	[-0.74,3.75]	0.171875
	Recovery AUC	+0.21	[-1.83,2.46]	0.843750

Localized First@1 exceeds Global-Pairwise First@1 on all four datasets. The paired Local-minus-Global First deltas are approximately +3.69 pp on UCI-HAR, +2.14 pp on N-MNIST, +1.86 pp on DVS-Gesture, and +7.20 pp on SHD; UCI-HAR, N-MNIST, and SHD are positive in all ten paired seeds.

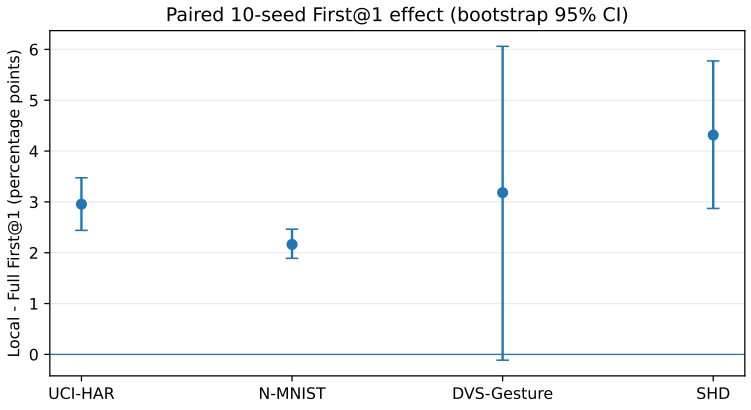

Figure 3: Paired Local-minus-Full First@1 effect with bootstrap 95% CIs.

B Exposure and capacity sensitivity

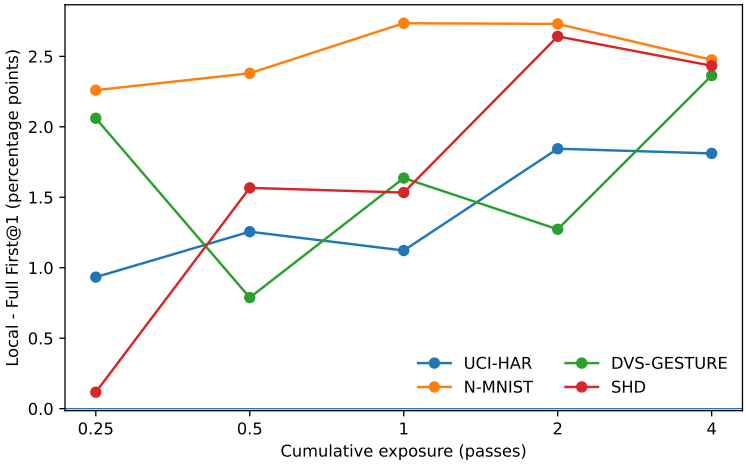

Figure 4: Audit Local-minus-Full First@1 advantage versus cumulative exposure. All 20 dataset-by-checkpoint differences are positive.

Table 4: Five-seed audit capacity sweep. Values are First@1 (%).

Dataset	8 slots	16 slots	32 slots	64 slots
UCI-HAR	88.29	89.51	89.42	88.62
N-MNIST	92.03	92.40	92.26	92.35
DVS-Gesture	48.00	48.61	47.76	46.30
SHD	57.90	57.78	57.78	57.98

Later DVS prefix representations reverse part of the apparent small-capacity preference. The supported conclusion is therefore that memory granularity interacts with addressing geometry, not that one capacity is universally optimal.

C Causal retry controls and memory ablations

Table 5: Representative five-seed retry controls (First@1 / Solve@5).

Policy	UCI-HAR	N-MNIST	DVS-Gesture
Full dynamic	77.43/98.64	85.03/98.99	39.32/86.21
No attempted exclusion	77.43/82.64	85.03/85.73	39.32/42.73
No-repeat random elimination	77.43/94.71	85.03/91.63	39.32/63.03
Retrieval off	28.73/96.91	8.65/57.93	10.61/56.59
Memory-action shuffle	9.92/88.33	19.10/58.74	10.45/52.88
Random action lookup	20.07/90.05	10.49/52.67	9.77/53.26

SHD is omitted from these earlier mechanism controls because they used the subsequently rejected semantic key; the final SHD closeout instead uses `temporal4x2048`. We do not treat old-key SHD ablations as evidence about the final representation.

Table 6: Independently trained mechanism ablations at four passes (First@1 / Solve@5).

Mode	UCI-HAR	N-MNIST	DVS-Gesture
Full	77.43/98.64	85.03/98.99	39.32/86.21
No memory	19.82/84.77	8.65/50.59	10.61/47.58
Positive only	80.35/99.34	64.42/70.10	31.14/65.61
Negative only	58.69/96.16	60.31/90.89	19.32/59.32
Bipolar evidence only	77.76/98.64	84.92/98.93	39.70/83.86
Unfiltered evidence only	21.59/83.56	10.54/50.73	9.92/45.45

D MATCHED AND END-TO-END REFERENCE SYSTEMS

Table 7 compares alternative learning rules on the same frozen representations: a linear BP readout, a two-layer MLP trained by BP, a local forward-style rule, and a reward-modulated Hebbian proxy. Table 8 instead gives fully supervised end-to-end references (MLP, surrogate-gradient LIF/BPTT, and temporal BP). Because the latter receive full class labels and use different update rules, both tables are contextual references rather than protocol-matched claims of superiority.

Table 7: Five-seed frozen-representation references at four passes: First@1 / Solve@5 (%).

Baseline	UCI-HAR	N-MNIST	DVS-Gesture	SHD
Linear BP	51.90/99.04	79.39/98.33	25.23/78.18	6.94/31.80
2-layer MLP BP	73.60/99.36	86.51/99.19	40.38/91.29	11.42/48.72
Local FF-style	37.73/96.76	23.36/69.94	8.03/39.32	4.84/24.29
Reward-modulated Hebbian proxy	63.27/100.00	77.81/98.50	20.91/71.29	5.11/25.80

Table 8: Five-seed end-to-end supervised references at four passes: First@1 / Solve@5 (%). These receive full class labels and are contextual rather than supervision-matched comparisons.

Baseline	UCI-HAR	N-MNIST	DVS-Gesture	SHD
Mean MLP BP	65.08/99.25	95.53/99.85	65.83/96.89	40.37/83.34
Surrogate LIF BPTT	78.91/99.93	95.91/99.94	57.12/98.03	51.46/87.62
Temporal BP	90.93/99.93	88.92/98.26	60.23/99.02	76.01/99.00

E REPRESENTATION DIAGNOSTICS

On N-MNIST, moving from the earlier 128-D random-projection key (RP128) to full2048_raw increases the Full First score from about 85.04% to 90.56%; a supervised MLP probe on the same frozen representation changes from about 84.73% to 91.71%. On UCI-HAR, full2048_raw improves Full First by about 2.04 pp over RP128. RP128 here refers to the earlier UCI/N-MNIST projection and is distinct from the DVS-specific `semantic128` closeout key. The probe is used only to test whether class-relevant information is extractable; it is not an upper bound on the memory architecture.

The five-view mean-probability ensemble reaches 71.64% First on the audit split. A closed-form quadratic recovery readout fused with the ensemble reaches 73.21% First and 98.06% Solve@5. Since union Top-2 and Top-3 coverage are 97.94% and 99.52%, the residual problem is primarily candidate discrimination.

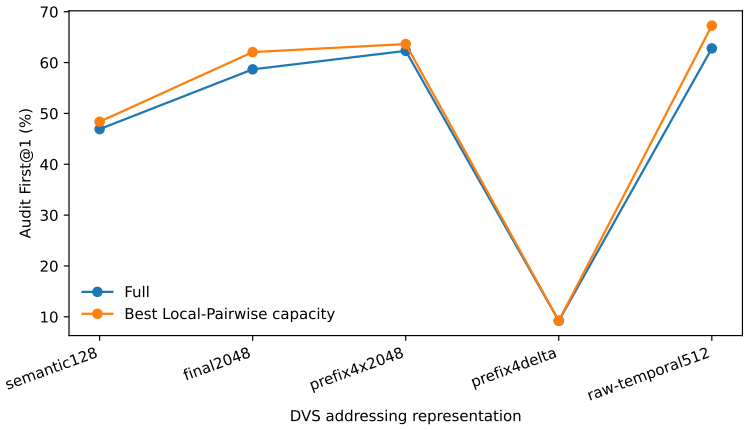

Figure 5: DVS audit representation ladder. The prefix-difference counterexample separates global class information from local episodic usability.

Table 9: DVS five-seed representation probes.

Representation	Dim.	Centroid top-1	1-NN	Purity@5
semantic128	128	48.00	40.61	33.87
final2048_raw	2048	49.70	48.73	40.07
prefix4x2048	8192	54.30	55.64	41.33
prefix4delta2048	8192	49.94	29.33	26.06
raw_temporal512	512	57.82	61.70	47.95

F BEST-ACHIEVED AUDIT EXTENSIONS

These values are mechanism-search headroom and must not be substituted for the frozen official closeout in confirmatory claims.

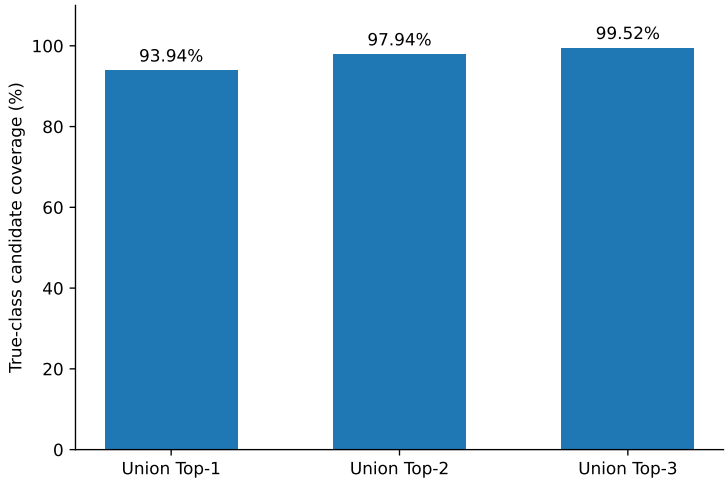

Figure 6: True-class coverage in the union of five DVS view candidate sets.

Table 10: Best achieved five-seed audit configurations after the frozen closeout. Here lowercase k is the audit-selected retrieval-neighbor count, distinct from the external attempt budget $K = 5$.

Dataset	First@1	Mechanism
UCI-HAR	94.24%	success posterior, retrieval $k = 4$, candidate Top-5, slots 32
N-MNIST	93.46%	success posterior, retrieval $k = 8$, candidate Top-3, slots 8
DVS-Gesture	73.21%	quadratic analytic + five-view ensemble
SHD	62.37%	local pair, retrieval $k = 32$, candidate Top-2, slots 16

G First-error and internal-dynamics diagnostics

At UCI-HAR pass 4, approximately 75.7% of First errors still place the truth in the top two, and exclusion-only rescue is about 58.6% among First errors. At pass 8, tanh saturation near ties is essentially zero, recurrent truth-to-wrong inversion is about 2.9%, and component cancellation is observed in about 75.3% of First errors. These diagnostics motivate candidate-resolution work but do not identify a single causal repair.

H Source/protocol provenance and reproducibility details

The reconstruction audit passes exact frozen hashes for the recovered core/evidence sources, compilation/import checks, absence of gradient-learning calls in reconstructed extensions, label-free prediction and trajectory-write signatures, and action emission before outcome comparison and persistent writes. The earliest historical `p1_trajectory_preference.py` source was not recovered byte-for-byte; the current module is therefore documented as a compatibility reconstruction of a confirmed interface/protocol rather than historical identity.

Main closeout details: seeds 7,17,27,37,47,57,67,77,87,97; four complete passes over the learn split; external action budget $K = 5$; one frozen official-test evaluation after pass 4 with no persistent cross-test writes; audit checkpoints at 0.25,0.5,1,2,4 passes; audit-only capacity sweep with 5 seeds and 8/16/32/64 slots per winner; representations UCI/N-MNIST full2048_raw, DVS semantic128, SHD temporal4x2048; paired bootstrap 95% CIs and exact two-sided sign-flip tests.