BEYOND RETENTION: CAUSAL LOCALIZATION OF POSITIONAL INTERFERENCE IN MEMORY-AUGMENTED VISION-LANGUAGE-ACTION POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory-augmented vision-language-action (VLA) policies can change their action predictions when historical evidence moves, even when retained content is matched. We distinguish evidence retention, upstream temporal encoding, downstream serialized memory-key positions, and numerical execution order in RoboMME. Canonical permutation controls establish exact recovery when represented memory and its key-position identity move together for all 27 primary episodes in both FrameSamp-Modul and TokenDrop-Modul. We then execute a bridge that preserves each retention-matched history's upstream representation while aligning its downstream key positions to a FAR anchor. Across 27 episodes evaluated at three inference seeds, FAR alignment reduces action divergence by 0.002948 MAE (95% task-stratified bootstrap CI $[0.001999, 0.004049]$) and exceeds common- and mismatched-shuffle controls, although task-specific effects are heterogeneous and negative for VideoPlaceOrder. Runtime instrumentation verifies unchanged initial queries and pre-RoPE memory keys while 432/512 key-position assignments change before RoPE, altering rotated keys and predicted actions. Neither a historical 120-rollout upstream correction nor a new 45-rollout downstream age-position intervention demonstrates improved task success. The latter activates on all 293 applicable policy calls but yields no paired rescues or regressions. These results connect downstream positional identity to the original action-level interference while separating causal localization from practical utility and independent-architecture generalization.

1 INTRODUCTION

Memory-augmented vision-language-action (VLA) policies use historical observations to support long-horizon robotic manipulation (Dai et al., 2026; Shi et al., 2026). However, retaining task-relevant evidence does not guarantee that a policy will use it consistently.

Prior work has established that robotic memory must support action-relevant retrieval and that historical content and temporal order can affect policy outputs (Guo et al., 2026; Yang et al., 2026; Rathi, 2026). We ask a narrower mechanistic question: when represented memory is held fixed, can reassigning its serialized key positions change the predicted action?

We begin with controlled FAR/MIDDLE/RECENT histories across 30 RoboMME episodes from three task families. Changing the temporal placement of relevant evidence alters initial action predictions. Retention-matched counterfactuals then show that fixing the identities of retained frames does not consistently eliminate this divergence. A historical geometric bridge based on separately executed upstream interventions does not support a simple mediation account, motivating a bridge at the downstream interface actually isolated below.

We identify two positional interfaces in the tested memory-as-modulation architecture: upstream per-frame temporal encoding and downstream serialized memory-key positions. To isolate the latter, we introduce a controlled U/V/W intervention. U uses the original represented memory and key positions; V permutes represented memory and assigns new serialized positions; and W applies the same permutation as V while transporting the original key positions.

The original three-unit diagnostic achieved strict U/W recovery in only one unit, exposing sensitivity to floating-point reduction order. Under a subsequently frozen canonical matched-reduction-order control, W recovers the reference action bitwise across all 27 primary FrameSamp episodes and three fresh process blocks, while V exceeds the predeclared displacement criterion in every corresponding check. The result also reproduces in TokenDrop, which changes the upstream perceptual-memory representation while retaining the same downstream modulation interface.

We then execute the previously missing bridge. For each retention-matched FAR/MIDDLE/RECENT history, we preserve its upstream representation and align only downstream memory-key positions to the FAR reference. Across 27 episodes and three inference seeds, FAR alignment reduces action divergence overall and relative to two shuffle controls, although the effect reverses for VideoPlaceOrder. Direct trained-checkpoint instrumentation verifies that the intervention reaches the intended pre-RoPE interface.

Finally, we separate mechanism from utility using two closed-loop studies. A historical 120-rollout experiment does not confirm improvement from an upstream temporal correction. A new 45-rollout development evaluation tests an online downstream age-position rule: it activates on all 293 applicable policy calls, but native, identity-control, and intervention arms all succeed on the same 7/15 paired episodes. Thus, the downstream intervention is operational, but no task-success benefit is demonstrated.

Our contributions are: (i) separating retained-evidence availability from residual temporal-history action sensitivity; (ii) causally localizing a reproducible action-level effect to serialized memory-key-position assignment across two upstream memory representations; (iii) bridging that interface to the original FAR/MIDDLE/RECENT residual under task-stratified controls; and (iv) distinguishing action-level mechanism from two negative closed-loop interventions, including an online downstream development study.

2 Related Work

Robotic memory and interference. RoboMME benchmarks memory-dependent robotic manipulation across multiple memory-augmented VLA variants (Dai et al., 2026). MemoryVLA, LongVLA, MemER, and HALO explore complementary mechanisms for representing, selecting, and using historical information (Shi et al., 2026; Fan et al., 2025; Sridhar et al., 2026; Shah et al., 2026). RoboMME-Interference demonstrates degradation under unrelated historical sessions and evaluates retrieval as a mitigation (Rathi, 2026). Chameleon goes beyond memory availability: its control-indexed policy retrieves actionable traces, and controlled trace-swapping and masking experiments test whether decisions follow relevant remembered information (Guo et al., 2026). Thus, the general distinction between retrieval and action-relevant memory use is established in prior work.

Temporal order and positional sensitivity. TemporalFlow-VLA learns physically grounded execution-history representations and uses controlled history-removal and shuffling tests to examine action sensitivity to historical content and temporal order (Yang et al., 2026). Position-dependent behavior has also been studied in long-context language models, including lost-in-the-middle effects and mechanistic analyses of positional bias (Liu et al., 2024; Wang et al., 2025). Our contribution is therefore not the observation that history order or position matters, but the isolation of serialized memory-key-position assignment at a specific downstream VLA interface.

Causal localization. Controlled internal interventions can distinguish the contributions of model components to predictions, although conclusions depend on intervention construction and evaluation metrics (Meng et al., 2022; Zhang & Nanda, 2024). We first separate retained-frame identity from residual temporal-history action sensitivity. We then hold represented memory fixed across a permutation and compare reassigned versus transported serialized key positions under matched numerical reduction order. The resulting U/V/W test isolates an action-level positional effect in the evaluated interface; it does not establish mediation of every temporal-history effect or a closed-loop performance benefit.

3 Methodology and Experimental Design

3.1 Memory-Augmented VLA Setting

We study the FrameSamp-Modul variant of the RoboMME memory-augmented VLA suite (Dai et al., 2026), using checkpoint `Yinpei/perceptual-framesamp-modul/79999`. Unless otherwise stated, inference uses seed 42. Our analysis focuses on the initial predicted action chunk under controlled interventions to historical visual memory.

FrameSamp-Modul uses a 512-token perceptual-memory budget, with 16 memory tokens per retained frame. We use the controlled history construction and fixed evaluation cohort specified in the frozen protocol. For the canonical intervention, represented memory is constructed from the FAR history and then manipulated at the downstream memory-attention interface. The current query observation and task instruction remain fixed across U/V/W conditions.

3.2 Memory Pathway and Positional Interfaces

The tested FrameSamp-Modul policy contains two positional interfaces relevant to our experiments. First, historical visual observations are selected and encoded into perceptual memory. During this upstream stage, FrameSamp uses a 768-dimensional PosEmb3D representation that includes a 256-dimensional temporal component. Changing this temporal information can therefore change the represented memory before downstream memory attention.

The represented memory is subsequently used by the MemoryAttention module. The current policy representation provides attention queries Q, while represented memory provides keys K and values V. Rotary positional embeddings (RoPE) are applied to the queries and memory keys, but not to the memory values. Consequently, changing the serialized position assigned to a memory key can change how an otherwise fixed represented-memory unit interacts with the current query.

These two interfaces should be distinguished. Upstream temporal-position interventions can change the represented memory itself, whereas the downstream U/V/W intervention used in this work changes serialized key-position identity while holding represented memory fixed.

3.3 Separating Memory Content from Positional Identity

Our central intervention distinguishes represented memory from the serialized positional identity assigned to that memory at the memory-attention interface. For FrameSamp, we use a fixed permutation of the 32 retained historical frame groups and compare three canonical conditions.

- U_c **(reference):** the canonical reference computation, retaining the original represented-memory arrangement and positional assignment.
- V_c **(reassigned position):** the represented memory is permuted, and positional identities follow the new serialized locations.
- W_c **(transported position):** the same represented-memory permutation as in V_c is used, but the original positional identity associated with each memory unit is transported with that unit.

The comparison between V_c and W_c therefore holds the represented-memory permutation fixed while changing whether positional identity is reassigned or transported. Recovery of U_c under W_c, together with displacement under V_c, provides evidence that the action difference depends on serialized positional identity rather than on memory permutation alone.

4 From Temporal Interference to Retention-Matched Controls

4.1 Controlled FAR/MIDDLE/RECENT Intervention

We first test whether temporal placement of task-relevant history changes the policy's initial action prediction. For each selected episode, we construct three histories in which semantically matched relevant evidence appears in FAR, MIDDLE, or RECENT positions while holding the query, task

family, distractor structure, memory budget, checkpoint, and seed fixed. The endpoint is the initial inferred 16×8 action chunk.

Across 30 scientific episodes from VideoUnmask, InsertPeg, and VideoPlaceOrder, all 540 planned directed process-level contrasts showed nonzero initial-action displacement. All 270 identity and replay control rows were exact, supporting intervention and replay integrity within the tested execution configuration. These findings establish initial-action sensitivity to temporal placement in this fixed cohort; the three process runs are not independent episodes. Figure 1 shows a concrete example from InsertPeg episode 22, in which the same retained memory identities are assigned FAR, MIDDLE, or RECENT temporal positions and the resulting initial action prediction changes. This example is illustrative; the cohort-level result is the full 30-episode analysis above. The complete six directed temporal comparisons are reported in Appendix B.

However, this first result does not by itself identify the mechanism. Moving evidence in the history can change which frames are retained by the frame-sampling memory system. A temporal-position effect could therefore arise because different evidence is available downstream, because the same evidence is represented differently downstream, or because of both. We therefore next ask whether the action divergence remains when the retained historical frames are explicitly matched.

4.2 Retention-Matched Counterfactuals

To test whether temporal-position sensitivity arises from changes in retained evidence, we compare natural and retention-matched predictions across the same 30 episodes. Natural histories use FrameSamp's ordinary frame selection; matched histories force the same retained-frame identities across FAR, MIDDLE, and RECENT. We analyze the initial 16×8 action chunks using preserved action arrays and frame-selection traces.

Each episode contributes three dependent temporal pairs, yielding 90 comparisons. We define the episode-level reduction as $\Delta d = d_{\text{natural}} - d_{\text{matched}}$: positive values indicate reduced action divergence after matching, and negative values indicate increased divergence. Figure 2 summarizes the median episode-level reduction for each task-family and temporal-pair combination.

We evaluate nine planned family-by-pair comparisons ($n = 10$ episodes each) using paired Wilcoxon signed-rank tests (Wilcoxon, 1945), with Holm correction (Holm, 1979). Table 1 reports natural-minus-matched group-median action MAE together with Holm-adjusted paired Wilcoxon p-values. The descriptive values discussed below and visualized in Figure 2 are instead medians of the episode-level paired reductions Δd.

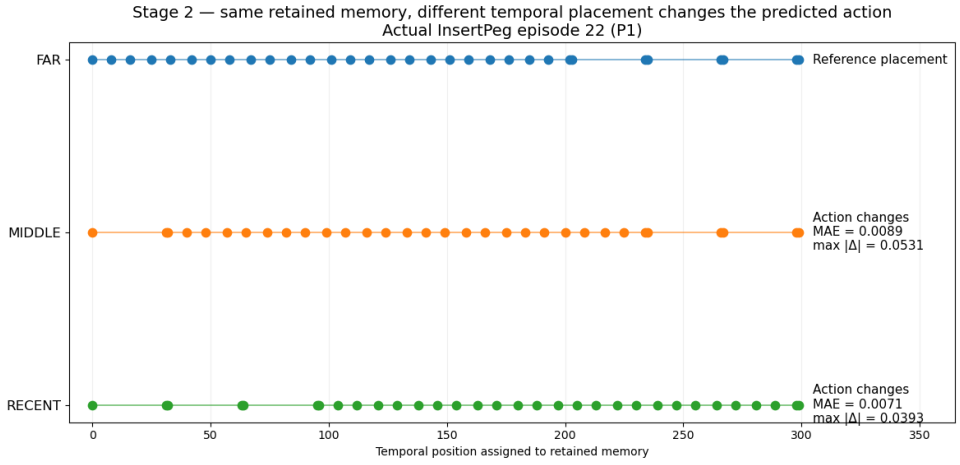

Figure 1: Concrete Stage-2 example from InsertPeg episode 22. The same retained memory identities are assigned different temporal placements under FAR, MIDDLE, and RECENT conditions, producing changes in the initial predicted action relative to the FAR reference. This episode is illustrative; the cohort-level result remains the full 30-episode analysis in which all 540 directed process-level contrasts showed nonzero initial-action displacement.

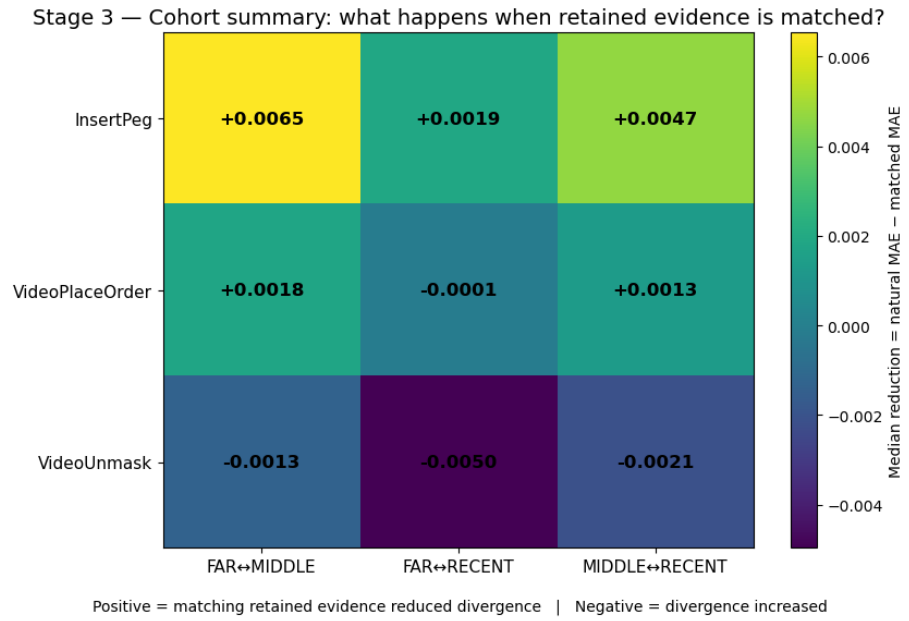

Figure 2: Retention-matching effect across task families and temporal pairs. Each cell shows the median of the 10 episode-level paired reductions, $\Delta d = d_{\text{natural}} - d_{\text{matched}}$. Positive values indicate that matching retained evidence reduced action divergence, whereas negative values indicate increased divergence after matching. Each cell represents 10 scientific episodes.

Retention matching has heterogeneous effects. For VideoUnmask, divergence increases after matching for all three temporal pairs, with median reductions of -0.00133, -0.00496, and -0.00209 for FAR–MIDDLE, FAR–RECENT, and MIDDLE–RECENT, respectively. The first two comparisons remain significant after Holm correction ($p = 0.03125$ and $p = 0.01758$). InsertPeg instead shows positive median reductions of 0.00653, 0.00188, and 0.00465. VideoPlaceOrder has mixed effects, including 0.00181 for FAR–MIDDLE and -0.00012 for FAR–RECENT.

Thus, retained-frame differences cannot consistently explain temporal-history action sensitivity in this cohort. This motivates the downstream key-position intervention in Section 5, without establishing that this interface mediates the original FAR/MIDDLE/RECENT effects.

Table 1: Retention-matching effects across nine family-by-pair comparisons. Entries report natural-minus-matched group-median action MAE; parentheses give Holm-adjusted paired Wilcoxon p-values ($n = 10$ episodes per family).

Task family	FAR–MIDDLE	FAR–RECENT	MIDDLE–RECENT
VideoUnmask	-0.001296 (0.03125)	-0.003701 (0.01758)	-0.001947 (0.09570)
InsertPeg	$+0.008736$ (0.19531)	$+0.002313$ (1.00000)	$+0.003685$ (0.48047)
VideoPlaceOrder	$+0.002176$ (0.16406)	$+0.002194$ (1.00000)	$+0.002247$ (0.18555)

4.3 Historical Upstream Geometric Bridge

We also evaluate whether the retention-matched action displacement directly points from one temporal condition toward another. For each directed comparison, we project the observed retention-matching displacement onto the corresponding upstream temporal-intervention direction. This historical geometric bridge is a stricter test than asking whether matching reduces pairwise distance:

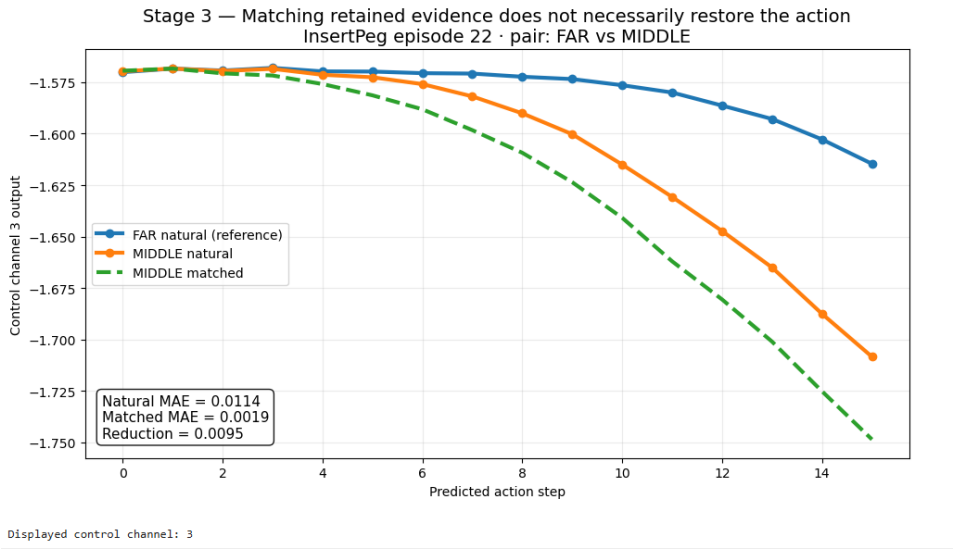

Figure 3: Concrete Stage-3 retention-matching example from InsertPeg episode 22. The plotted control channel compares the FAR natural reference, the MIDDLE natural prediction, and the MIDDLE retention-matched prediction. In this episode, matching retained evidence substantially reduces the action divergence. Across the full cohort, however, the effect is heterogeneous, so retention matching is not a sufficient general explanation of the temporal-history sensitivity.

it asks whether the retention-matching change moves the action output in the direction needed to explain the temporal-position effect.

The recovered historical bridge analysis contains 180 directed records, averaging the three nuisance-process values within each episode and directed comparison. Overall, the projection fraction has median -0.031, the cosine alignment has median -0.461, and the median distance reduction is approximately -4.0×10^{-4}. Only 48.3% of directed projections are positive. Thus, this upstream bridge does not support a simple geometric account in which retention matching consistently moves the action toward the alternative temporal condition.

Together, the retention-matched and historical bridge analyses suggest that availability of relevant frames is only one part of the action-sensitivity phenomenon. The next section therefore intervenes on the downstream positional identity assigned to represented memory. This intervention tests a causal positional interface; it should not be read as a complete mediation analysis of all temporal-history effects.

5 Causal Localization of Serialized Memory Position

5.1 Temporal and Sequence-Order Decomposition

The controlled temporal-position and sequence-order analyses revealed context-dependent, non-additive action effects across 90 dependent episode-pair comparisons. The median interaction-to-mean-temporal L_2 ratio was 1.868, while the median temporal-row cosine was -0.395. A separate sequence-order isolation analysis yielded a median sequence/native distance ratio of 1.100. The complete descriptive statistics and experimental definitions are provided in Appendix C, Table 2.

5.2 U/V/W Intervention

We test whether serialized key-position assignment causally affects action predictions in downstream RoPE-based memory attention (Su et al., 2024). For each episode, we fix the current query and represented memory and compare three conditions under a common memory-unit permutation π.

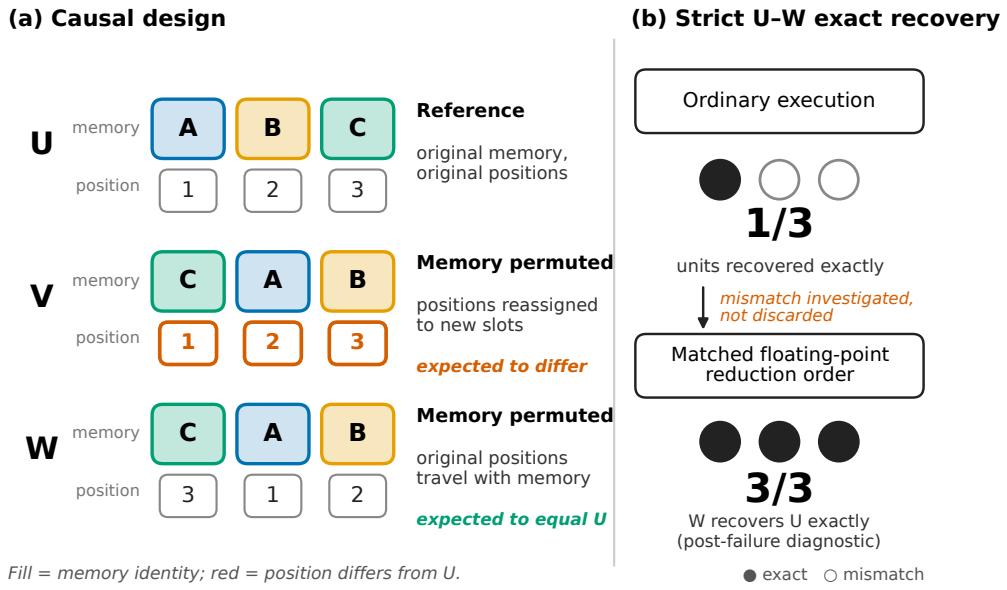

Figure 4: U/V/W causal design and numerical-control diagnostic. Under ordinary execution, strict U–W recovery held in only 1/3 selected diagnostic units. A subsequent same-process investigation identified sensitivity to floating-point reduction order. After matching the reduction order, W recovered U exactly in all 3/3 diagnostic units. This post-failure diagnostic motivated the prospectively frozen canonical protocol and does not replace the original 1/3 result.

Let m_i denote a represented memory unit with original key position p_i. The reference condition U preserves (m_i, p_i). V permutes the memory but assigns new serialized slot positions $(m_{\pi(i)}, p_i)$, whereas W applies the same permutation while transporting the original positions $(m_{\pi(i)}, p_{\pi(i)})$.

V and W therefore contain identical permuted memory; only their positional assignments differ. In exact arithmetic, jointly transporting memory, positions, and the corresponding attention mask makes W invariant to the permutation. Recovery of U under W, together with displacement under V, therefore isolates the effect of serialized key-position assignment under controlled numerical execution. We test U/W bitwise recovery on the full raw 20×8 output and U/V displacement on the initial 16×8 action chunk. The former is the stricter equality criterion.

5.3 Original Diagnostic and Numerical Control

The initial three-unit diagnostic exposed a numerical confound: strict U/W recovery under ordinary execution held in only one unit. Follow-up same-process tests identified floating-point reduction-order sensitivity in the two failures (see Appendix D).

We therefore froze a canonical matched-reduction-order protocol before the full-cohort evaluation. This protocol uses a common reduction order across U, V, and W while preserving the scientific contrast: V receives newly serialized key positions, whereas W transports each memory unit's original positions. The checkpoint, query, represented memory, and action endpoint remain unchanged. The original $1/3$ diagnostic remains distinct from the prospective results reported below. Figure 4 summarizes the U/V/W causal design and the numerical-control sequence that motivated the canonical matched-reduction-order protocol used in the prospective full-cohort evaluation.

5.4 Prospective Full-Cohort FrameSamp Result

We then ran a prospective full-cohort FrameSamp experiment under the canonical control. The primary scientific cohort contains 27 episodes. Three additional episode-0 units are reserved as engineering controls and are excluded from the primary sample. Each primary unit is evaluated in three fresh process runs, P1, P2, and P3.

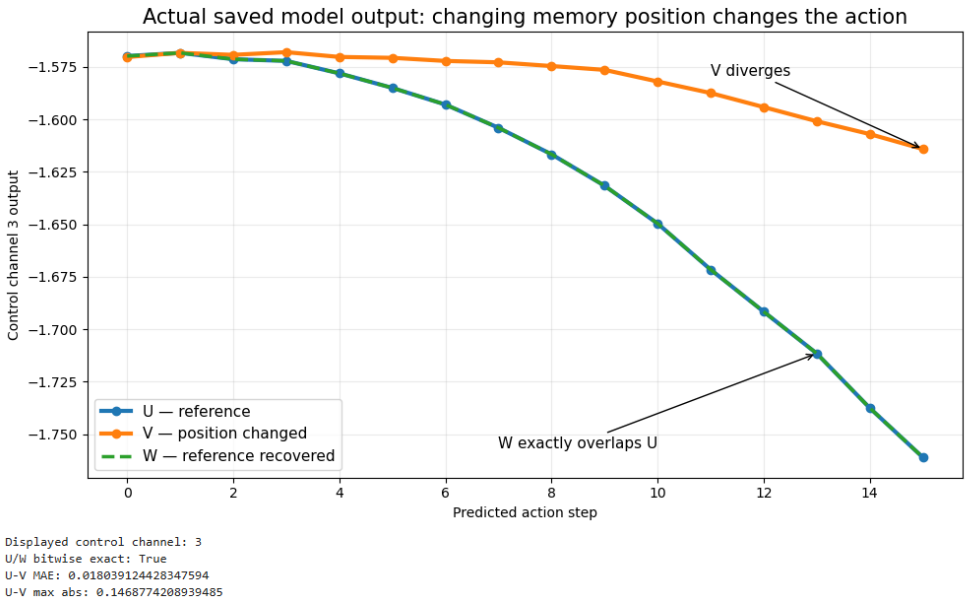

Figure 5: Concrete canonical U/V/W intervention on a saved FrameSamp model output from Insert-Peg episode 22. U_c is the reference prediction. V_c uses the same represented-memory permutation as W_c but reassigns serialized key positions to the new slots, producing a different predicted action. W_c transports the original key-position identities with the same represented-memory permutation and exactly recovers U_c. The plotted trace is illustrative; the scientific result is the frozen 27-episode cohort, for which 27/27 primary episodes passed the joint criterion, with 81/81 displaced U_c–V_c checks and 81/81 exact U_c–W_c checks across three fresh process runs.

All 27 primary episodes pass the joint causal gate. Across the three process runs, W exactly recovers U at the raw 20×8 action level in all $81/81$ process-level checks. In the same checks, V remains displaced from U on the analyzed 16×8 action chunk in all $81/81$ cases, exceeding the frozen legacy nuisance bounds. Thus, when represented memory content and reduction order are controlled, transporting the original serialized key positions is sufficient to recover the reference action, whereas assigning the new serialized key positions is sufficient to preserve the action displacement. Figure 5 shows a concrete saved-output trace from the canonical intervention. The cohort-level result is reported directly in the text and in Appendix E.

The appropriate scientific denominator is the fixed cohort of 27 primary episodes, not 81 independent samples. The 81 process-level checks are replication and integrity checks over three fresh runs of the same fixed episodes. We therefore report this experiment as a fixed-cohort census rather than as random-population inference, and we do not attach a population-level p-value to the mechanistic gate.

5.5 Executed Bridge to the Original Temporal Residual

The U/V/W result establishes that downstream key-position assignment can change actions for fixed represented memory, but does not by itself show that this interface contributes to the original FAR/MIDDLE/RECENT residual. We therefore execute a cross-experiment bridge on the retention-matched histories. For every matched layout, we preserve the selected frames and their upstream representation, then map each semantic frame identity to its rank in the FAR anchor and expand that rank into 16 within-frame token coordinates. Only the downstream memory-key positions are changed. Native-to-canonical replay, deterministic replay, a common random identity map, and condition-specific shuffles control for numerical execution and nonspecific remapping.

Across 27 fixed episodes evaluated at inference seeds 42, 7, and 2026, FAR alignment reduces action divergence by 0.002948 MAE (95% task-stratified episode-bootstrap CI $[0.001999, 0.004049]$). The reduction exceeds a common shuffle by 0.001300 ($[0.000663, 0.001917]$) and condition-specific

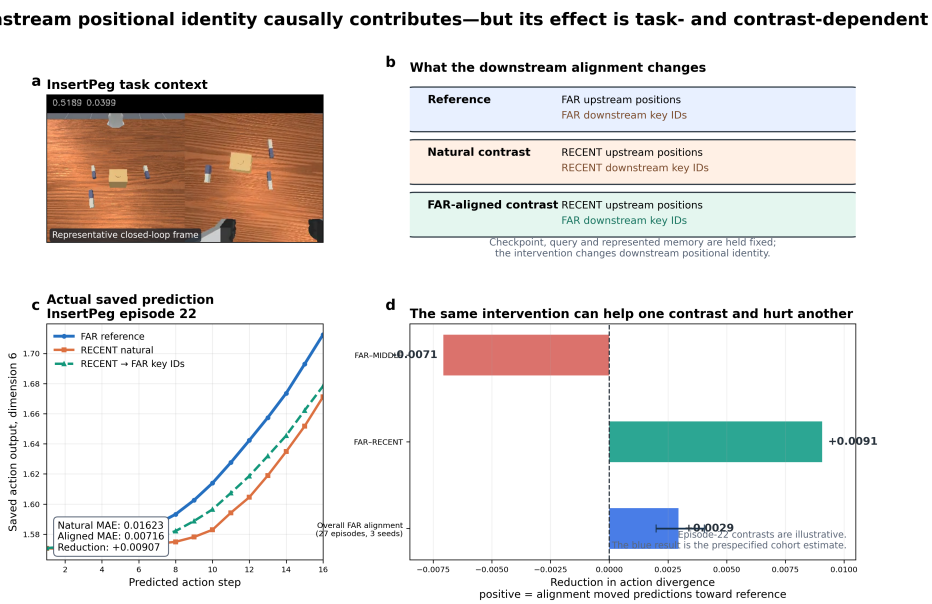

Figure 6: Executed downstream bridge. The FAR-aligned condition preserves each matched history's selected frames and upstream representation while assigning downstream memory-key positions from the FAR anchor. The saved InsertPeg episode-22 action trace illustrates the intervention; the task-specific and overall estimates use the prespecified 27-episode, three-seed cohort. Positive values indicate reduced divergence from the FAR reference.

mismatched shuffles by 0.002089 ($[0.001377, 0.002835]$). The effect is heterogeneous: VideoUnmask improves by 0.006965, InsertPeg by 0.002913, and VideoPlaceOrder changes by -0.001035. Seeds are repeated inference evaluations of the same episodes, not independent scientific samples. The negative VideoPlaceOrder estimate rules out a universal correction.

5.6 Trained-Checkpoint Interface Verification

Runtime instrumentation captures tensors immediately before RoPE in the trained checkpoint. On the instrumented matched input, native and aligned conditions have bitwise-identical initial queries, query positions, memory cache, and unrotated memory keys. FAR alignment changes 432 of 512 downstream key-position assignments; the rotated keys change and the predicted-action MAE is 0.004952. This verifies that the bridge reaches the intended interface rather than altering frame selection or upstream content.

5.7 Interpretation

Jointly permuting represented memory, its original key-position identities, and the corresponding mask is an expected attention invariance, not itself a novel mechanism. Under canonical matched reduction order, W recovered U bitwise in all 81 process-level checks over 27 fixed primary episodes.

The decisive contrast is between V and W: both use the same permuted represented memory, but V reassigns key positions to the new serialized slots, whereas W transports their original identities. V exceeded the predeclared numerical nuisance criterion in all 81 checks. Thus, serialized key-position assignment causally affects initial action predictions when represented memory is held fixed in the tested FrameSamp-Modul interface. The executed bridge further shows that aligning this interface reduces the original FAR-anchored residual overall and relative to shuffle controls. The effect remains task dependent, and neither result establishes a closed-loop task-success benefit.

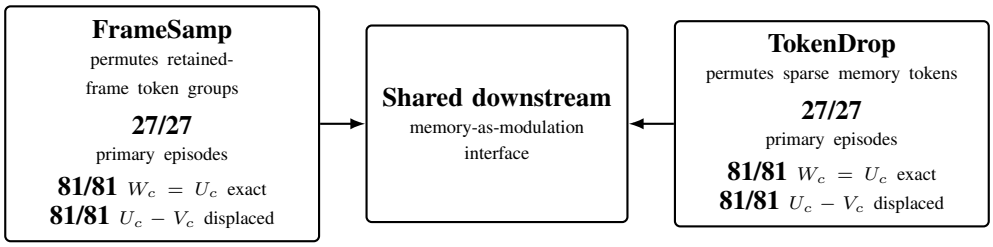

Figure 7: Cross-representation replication under a shared downstream memory-as-modulation interface. For both FrameSamp and TokenDrop, 27/27 primary episodes passed the frozen joint criterion; W_c recovered U_c in 81/81 checks and $U_c - V_c$ remained displaced in 81/81 checks, across three fresh process runs per representation. This is not replication across independent architecture families.

6 Cross-Representation Replication

6.1 Motivation

The FrameSamp experiment localizes action sensitivity to downstream serialized memory-key positions, but uses frame-sampled perceptual memory. We test whether this effect replicates under TokenDrop-Modul, which instead uses sparse token dropping before the shared downstream memory-as-modulation interface (Dai et al., 2026). This tests cross-representation replication, not generalization across independent VLA architectures.

6.2 TokenDrop Protocol

We use the same fixed 27-primary-episode cohort as the FrameSamp Stage-7 experiment. The same three episode-0 engineering controls are reserved for preflight validation and are excluded from the scientific sample. Each primary episode is evaluated in three fresh process runs, P1, P2, and P3.

The TokenDrop experiment uses the same U/V/W causal structure. U is the reference condition. V uses the represented memory permutation with newly serialized key positions. W uses the same represented memory permutation as V but transports the original serialized key positions with the represented memory. As in the FrameSamp experiment, the strict recovery gate requires W to match U at the raw 20×8 action level, while the displacement gate requires V to remain displaced from U on the analyzed 16×8 action chunk.

The prospective TokenDrop run includes an amendment to correct an engineering configuration issue observed in the initial D0 validation attempt. After this amendment, the D0 controls passed and the primary fixed-cohort evaluation was run under the amended frozen protocol. We treat the failed D0 attempt as an engineering preflight issue, not as a primary scientific observation. Figure 7 illustrates the two perceptual-memory representations and summarizes their positional-transport outcomes.

6.3 Replication Result

On the analyzed 16×8 action chunk, the median episode-level U_c–V_c MAE (interquartile range) was 0.002659 (0.001171–0.006344) for FrameSamp and 0.002171 (0.001258–0.003949) for TokenDrop. We averaged each episode's three process-run measurements before summarizing the 27 primary episodes. The corresponding median maximum absolute displacements were 0.022599 and 0.018321, respectively. TokenDrop selected 254–512 active tokens across these episodes under its 512-token budget. These action-displacement magnitudes do not establish a closed-loop performance benefit or rank the two representations.

All 27 primary TokenDrop episodes passed the frozen joint criterion. Across three fresh process runs, W_c recovered U_c bitwise on the raw 20×8 output in all $81/81$ checks, while V_c exceeded the predeclared displacement criterion on the analyzed 16×8 action chunk in all $81/81$ checks. The

scientific denominator remains 27 fixed episodes: the process runs are reproducibility checks, not independent scientific episodes.

6.4 Scope of the Replication

The canonical U/V/W result reproduces across FrameSamp and TokenDrop, two different upstream perceptual-memory representations sharing the same downstream memory-as-modulation interface. This supports the causal role of serialized key-position assignment across the tested representations, but does not establish that the effect generalizes to other VLA architectures, memory interfaces, or task distributions.

7 Closed-Loop Behavioral Evaluation

7.1 Motivation

The preceding experiments establish a causal effect of serialized key-position assignment on initial action prediction. Whether such action-level sensitivity translates into improved robotic task performance is a separate question: closed-loop outcomes also depend on feedback, controller dynamics, and task-specific failure modes. We therefore evaluate two distinct interventions. The historical prospective study changes upstream temporal representation. A new development study changes online downstream memory-key positions using only timestamps already observed by the policy. Neither study is used to redefine the mechanistic endpoint.

7.2 Historical Upstream Rollout Protocol

We evaluate 120 prospective rollouts across four conditions, with 30 rollouts per condition. The conditions are CLEAN, RECENT_NATIVE, FAR_NATIVE, and FAR_TPOS_RECENT. CLEAN is the ordinary non-intervened baseline for the selected rollout setting. RECENT_NATIVE and FAR_NATIVE evaluate native history placements. FAR_TPOS_RECENT tests whether modifying the positional treatment of FAR history toward the RECENT condition improves closed-loop behavior. The closed-loop protocol is not used to redefine the mechanistic result. It is a separate behavioral test of temporal-position correction under simulator feedback, not a direct closed-loop evaluation of the downstream U/V/W serialized-key-position transport intervention.

7.3 Historical Upstream Result

The prospective evaluation did not confirm the pre-specified behavioral improvement. Success counts were $8/30$ for FAR_NATIVE, $7/30$ for CLEAN, $5/30$ for FAR_TPOS_RECENT, and $4/30$ for RECENT_NATIVE. Relative to FAR_NATIVE, RECENT_NATIVE showed a -13.3 percentage-point difference (exact McNemar $p = 0.125$; Holm-adjusted $p = 0.25$), while FAR_TPOS_RECENT showed a -10.0 percentage-point difference (exact McNemar $p = 0.375$; Holm-adjusted $p = 0.375$). Neither comparison confirmed the hypothesized improvement. Importantly, this experiment tests an upstream temporal-position correction and is therefore distinct from the downstream serialized memory-key-position intervention isolated by the U/V/W experiments. The complete paired-test table is provided in Appendix H, Table 7. Figure 8 shows the observed success rates with 95% Wilson confidence intervals and outcome-count tiles.

7.4 Downstream Development Evaluation

We separately test whether an online downstream position rule improves closed-loop success. At policy step t, a frame acquired at step s receives coordinate $31 - \min(t - s, 31)$, expanded to its 16 memory-token coordinates. The rule uses no relevance annotation, future observation, ground-truth action, subgoal predictor, or success feedback. Fifteen paired task episodes from VideoUnmask, InsertPeg, and VideoPlaceOrder are evaluated under native, identity-control, and age-position arms, yielding 45 official RoboMME rollouts. This is a development cohort, not a held-out efficacy study.

Native and identity-control outcomes agree on all 15 episodes. The age-position intervention changes downstream positions on all 293 applicable age-arm policy calls, but every paired outcome

remains unchanged: native, identity-control, and intervention arms each succeed on 7/15 episodes, with zero rescues and zero regressions against either control. Thus, the intervention is operational and changes trained-policy actions, but no practical task-success benefit is demonstrated in this cohort.

7.5 Interpretation

Neither closed-loop study confirms a task-success improvement. The historical 120-rollout result concerns an upstream temporal-position correction. The new 45-rollout result establishes that downstream position changes can be applied during actual policy control, but the tested timestamp rule yields no paired rescue or regression. The privileged FAR-alignment map used for causal localization is not a deployable policy and is not evaluated as an efficacy claim. Mechanistic localization, operational intervention, and beneficial control therefore remain separate conclusions.

8 Discussion and Limitations

Our experiments distinguish memory availability from downstream positional utilization. Retention matching does not consistently remove temporal-history-induced action divergence, while the historical geometric bridge does not support a simple upstream mediation account. The U/V/W intervention establishes a more specific result: under canonical matched reduction order, serialized key-position assignment causally affects initial action predictions even when represented memory is held fixed. The executed downstream bridge then connects that interface to the original residual: FAR alignment reduces divergence overall and relative to shuffle controls, but the effect is task dependent.

This effect reproduces across FrameSamp and TokenDrop, which use different upstream memory representations but share the same downstream modulation interface. It does not establish that dis-

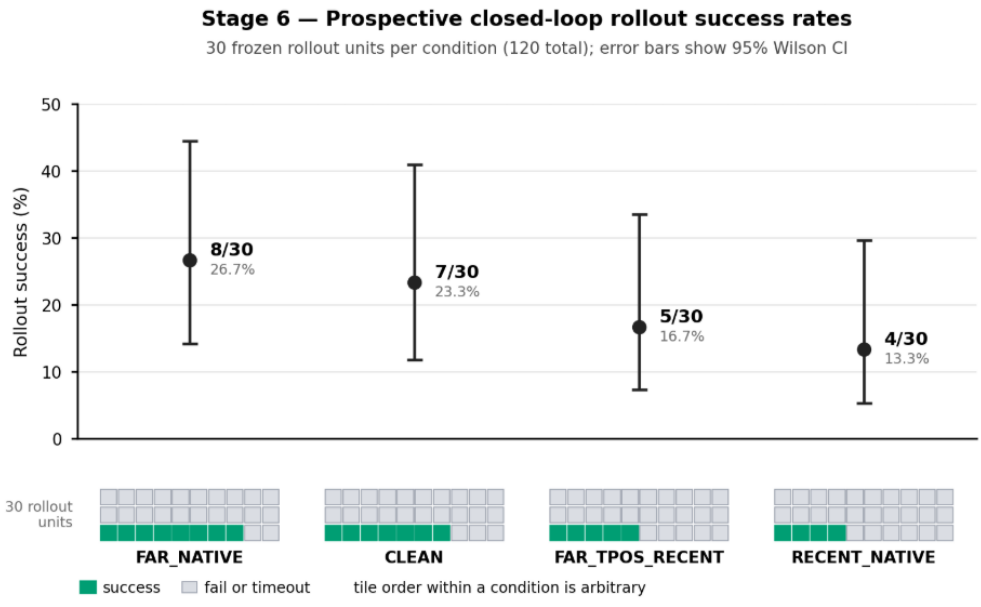

Figure 8: Prospective closed-loop rollout success rates across four conditions, with 30 frozen paired rollout units per condition (120 total rollouts). Points show observed success rates and error bars show 95% Wilson confidence intervals; the tiles visualize success versus non-success counts. FAR_NATIVE achieved 8/30 successes, CLEAN 7/30, FAR_TPOS_RECENT 5/30, and RECENT_NATIVE 4/30. The pre-specified behavioral improvement was not confirmed. This experiment evaluates an upstream temporal-position correction, not downstream U/V/W serialized-key-position transport.

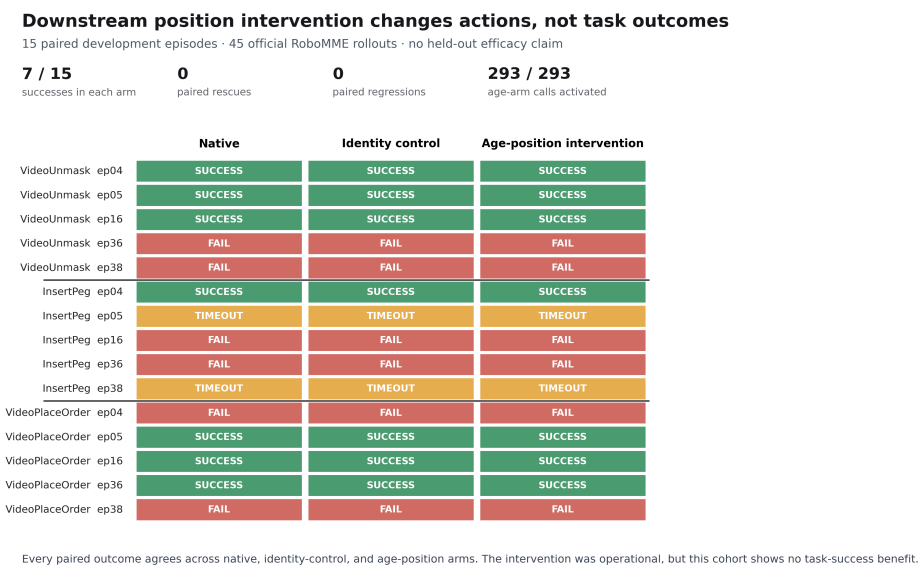

Figure 9: Downstream closed-loop development evaluation. Rows are paired task episodes and columns are native, identity-control, and online age-position arms. The intervention activates on 293/293 applicable calls, yet produces no paired task-success disagreement. This is an operational negative result, not a held-out efficacy estimate.

placed actions cause task failures. The two closed-loop studies strengthen this distinction. Neither the upstream correction nor the online downstream age-position rule improves task success in its tested cohort. The downstream rule is verified as operational on every applicable policy call, so its null paired outcome is not an intervention-delivery failure.

Limitations. The mechanistic results concern a fixed cohort of 27 episodes from three RoboMME task families and one shared downstream architecture. Process repeats establish reproducibility on that cohort, not population-level generalization. Exact recovery is conditional on canonical numerical reduction order; the original ordinary diagnostic passed only one of three selected units. The bridge uses privileged semantic identity to construct FAR alignment, whereas the deployable timestamp rule is a coarser intervention that can collapse old frames. The 45-rollout cohort is developmental rather than held out, and its absence of outcome disagreements cannot establish equivalence. Future work should test independent downstream memory architectures and prespecified deployable alignment rules on a larger held-out cohort.

9 Conclusion

We investigated whether temporal-history sensitivity in memory-augmented VLA policies can persist when task-relevant evidence is retained. Across 30 RoboMME episodes, moving relevant history changed initial action predictions, while matching retained frame identities did not consistently eliminate divergence. A subsequent controlled U/V/W intervention isolated a downstream action-level dependence on serialized memory-key positions: under canonical matched reduction order, transporting original positions recovered the reference action bitwise across all 27 primary FrameSamp episodes and three fresh process blocks, whereas reassigning positions produced displacement beyond the frozen nuisance criterion. The result reproduced in TokenDrop, a different upstream memory representation sharing the same downstream modulation interface.

An executed bridge then preserved each matched history's upstream representation while aligning downstream key positions to FAR. Across 27 episodes and three inference seeds, alignment reduced FAR-anchored action divergence overall and relative to shuffle controls, while exposing a negative VideoPlaceOrder effect that rules out a universal correction. Runtime instrumentation verified the intervention at the trained checkpoint's pre-RoPE interface.

Neither a historical 120-rollout upstream correction nor a new 45-rollout downstream development intervention improved official task success. The latter activated on all 293 applicable calls but produced no paired rescue or regression. Together, the experiments causally localize an action-level contribution of downstream positional identity while showing why causal explanation, deployable intervention, and behavioral benefit must be evaluated separately.

Reproducibility Statement

We preserve the frozen experimental protocols, fixed evaluation cohorts, executed code, process-level replication records, engineering controls, and raw result artifacts for the principal experiments. The appendix describes the intervention definitions, numerical controls, episode selection, and analysis procedures. We distinguish scientific episodes from repeated nuisance-process evaluations and preserve the original diagnostic failures and negative behavioral results. The added evidence package includes three-seed bridge action bundles, trained-checkpoint RoPE captures, 45 downstream rollout receipts, per-call intervention logs, videos, immutable manifests, and SHA-256 inventories.

AI Use Statement

Generative AI tools were used during this project for research brainstorming, hypothesis refinement, experimental-design feedback, implementation and debugging assistance, literature discovery, analysis and interpretation support, AI-assisted creation and revision of scientific figures and figure-generation code, and drafting and editing of manuscript text. The author remained responsible for experimental decisions, execution, verification of reported results, interpretation, citations, and the final manuscript.

Stage 2 directional summaries were checked against original episode-result logs. The principal retention-matching and U/V/W statistics were independently reconstructed from preserved raw action arrays. Historical geometric-bridge statistics were checked against preserved derived records, but were not independently reconstructed from all underlying temporal-intervention action arrays.

References

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. RoboMME: Benchmarking and understanding memory for robotic generalist policies. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.04639.

Yiguo Fan, Shuanghao Bai, Xinyang Tong, Pengxiang Ding, Yuyang Zhu, Hongchao Lu, Fengqi Dai, Wei Zhao, Yang Liu, Siteng Huang, Zhaoxin Fan, Badong Chen, and Donglin Wang. Long-VLA: Unleashing long-horizon capability of vision language action model for robot manipulation. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2018–2037, 2025. URL https://proceedings.mlr.press/v305/fan25a.html.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, 1991. doi: 10.1145/103162.103163.

Xinying Guo, Chenxi Jiang, Hyun Bin Kim, Yuhang Han, Ying Sun, Yang Xiao, and Jianfei Yang. Chameleon: Control-indexed prospective memory for visuomotor manipulation, 2026. URL https://arxiv.org/abs/2603.24576.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Soumil Rathi. Robomme-interference: Benchmarking robot memory under interference, 2026. URL https://arxiv.org/abs/2606.22338.

Rutav Shah, Yisu Li, Femi Bello, Yuke Zhu, and Roberto Martín-Martín. Memory retrieval in visuomotor policies for long-horizon robot control. In *Proceedings of Robotics: Science and Systems*, 2026. URL https://arxiv.org/abs/2606.25136.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=54U3XHf7qq.

Ajay Sridhar, Jennifer Pan, Satvik Sharma, and Chelsea Finn. Scaling up memory for robotic control via experience retrieval. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9da515b1ad19d032a7398f00f5ff9b0c-Abstract-Conference.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fvkElsJOsN.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Jiarui Yang, Yehao Lu, Yuning Su, Yu Zhong, Yufeng Xie, Yazhou Zhang, Haiyu Lan, Kaixiang Lu, Peiwen Lin, Chuang Wang, Junwei Liang, and Enyu Li. Temporalflow-vla: Learning physically grounded execution history for long-horizon robot manipulation, 2026. URL https://arxiv.org/abs/2608.26821.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

A Additional Experimental Details

A.1 Model and Evaluation Cohort

We evaluate the FrameSamp-Modul variant of RoboMME (Dai et al., 2026), using checkpoint Yinpei/perceptual-framesamp-modul/79999. Unless otherwise stated, inference uses seed 42. FrameSamp-Modul has a 512-token perceptual-memory budget, comprising 16 memory tokens per retained frame and 32 retained frames.

The initial evaluation comprises 30 fixed scientific episodes from three RoboMME task families: VideoUnmask, InsertPeg, and VideoPlaceOrder, with 10 episodes per family. All controlled comparisons use the corresponding episode's fixed task instruction and current query observation.

A.2 Experimental Units and Action Endpoints

The temporal-position experiment constructs FAR, MIDDLE, and RECENT histories for each episode. It evaluates six directed temporal comparisons per episode across three process blocks, yielding 540 process-level comparisons. These process-level measurements are repeated evaluations of the 30 scientific episodes, not 540 independent samples.

The retention-matched analysis considers the three undirected temporal pairs for each episode, yielding 90 dependent episode-pair comparisons and nine planned family-by-pair statistical tests. Each test contains 10 scientific episodes.

For these experiments, the primary action-comparison endpoint is the initial predicted 16×8 action chunk. The original raw action outputs have shape 20×8.

A.3 Prospective Mechanistic Cohort

The prospective FrameSamp U/V/W experiment uses 27 fixed primary episodes. Three additional episode-0 units serve as engineering controls and are excluded from the primary scientific cohort.

Each primary episode is evaluated in three fresh process blocks, P1, P2, and P3. Accordingly, 81 episode–process checks assess reproducibility on the same 27 scientific units; they do not constitute 81 independent episodes.

The strict U/W recovery criterion requires bitwise equality of the raw 20×8 action outputs under canonical matched reduction order. The U/V displacement criterion is evaluated on the initial 16×8 action chunk against the predeclared numerical nuisance criterion.

B Temporal-Position and Retention-Matching Results

B.1 Controlled Temporal-Position Intervention

For each of the 30 scientific episodes, we construct three controlled histories in which semantically matched task-relevant evidence is placed in FAR, MIDDLE, or RECENT temporal locations. The current query observation, task family, distractor structure, memory budget, checkpoint, and inference seed are held fixed.

We compare all six directed temporal transitions: FAR→MIDDLE, FAR→RECENT, MIDDLE→FAR, MIDDLE→RECENT, RECENT→FAR, and RECENT→MIDDLE.

Across the three process blocks, this yields 540 directed process-level comparisons. All 540 comparisons exhibit nonzero action displacement on the initial 16×8 predicted action chunk. These process-level measurements are repeated evaluations of the same 30 scientific episodes and are not treated as independent samples. All 270 identity and replay control rows were exact.

B.2 Retention-Matched Counterfactual

The retention-matched experiment separates changes in retained frame identity from downstream effects that persist after retention is controlled. For each episode and temporal condition, the natural arm uses the frames retained by the ordinary FrameSamp pipeline. The matched arm overrides frame selection so that retained historical frame identities are matched across the temporal conditions being compared.

For each episode, we analyze the three undirected temporal pairs: FAR–MIDDLE, FAR–RECENT, and MIDDLE–RECENT. This produces 90 dependent episode-pair comparisons across the 30 scientific episodes.

For a temporal pair, let d_{natural} denote the action distance between the two natural conditions and let d_{matched} denote the corresponding action distance after retention matching. We define the episode-level reduction as

$$\Delta d = d_{\mathrm{natural}} - d_{\mathrm{matched}}.$$

Positive values therefore indicate that matching retained frames reduces temporal-position divergence, whereas negative values indicate increased divergence after matching.

B.3 Statistical Analysis

The analysis contains nine pre-specified family-by-pair tests: three temporal pairs within each of the three task families. Each test contains 10 scientific episodes.

Paired Wilcoxon signed-rank tests are applied to the episode-level natural-versus-matched measurements (Wilcoxon, 1945). The resulting p-values are corrected jointly across the nine planned comparisons using Holm's sequential procedure (Holm, 1979).

Table 1 reports the difference between the natural and matched *group medians*. This quantity is distinct from the median of the 10 episode-level paired reductions Δd used for descriptive paired summaries. The two statistics should therefore not be interpreted interchangeably.

The Holm-adjusted paired tests are significant for two VideoUnmask comparisons: FAR–MIDDLE ($p_{\mathrm{Holm}} = 0.03125$) and FAR–RECENT ($p_{\mathrm{Holm}} = 0.01758$). The remaining seven planned comparisons do not cross the 0.05 adjusted threshold. These results support a heterogeneous retention-matching effect rather than a uniform retention-only explanation.

B.4 Historical Upstream Geometric Bridge

We additionally test whether the action displacement induced by retention matching geometrically aligns with the original temporal-position displacement. For each directed temporal comparison, the retention-matching displacement vector is projected onto the corresponding temporal-intervention direction.

The recovered analysis contains 180 directed records after averaging the three nuisance-process values within each episode and directed comparison. The median projection fraction is -0.031, median cosine alignment is -0.461, and median distance reduction is approximately -4.0×10^{-4}. Only 48.3% of directed projections are positive.

Thus, retention matching does not consistently move the action output toward the alternative temporal condition. The historical upstream bridge therefore does not support a simple mediation account in which retention differences alone explain the original FAR/MIDDLE/RECENT action effect.

C Temporal and Sequence-Order Decomposition

C.1 Two Positional Interfaces

The tested memory-as-modulation architecture contains two distinct positional interfaces.

The first is the upstream per-frame temporal representation. In the inspected FrameSamp implementation, a 768-dimensional PosEmb3D representation includes a 256-dimensional temporal component. The position embedding is projected and passed through a SiLU nonlinearity before being integrated into the encoded perceptual-memory representation. Changing this temporal information can therefore change the represented memory before downstream language-model memory attention.

The second interface assigns serialized positions to represented memory tokens during downstream memory attention. In the inspected MemoryAttention implementation, queries and memory keys receive rotary position embeddings (RoPE), whereas memory values do not (Su et al., 2024). The original memory keys normally receive positions according to their serialized token indices.

These interfaces should not be conflated. Altering the upstream temporal embedding can change the encoded memory representation. Altering downstream serialized key positions can change how an otherwise fixed represented memory is accessed during attention.

Table 2: Descriptive medians from the temporal-position and serialized sequence-order analyses. Each analysis contains 90 dependent episode-pair comparisons.

Statistic	Median
Temporal-row cosine	-0.395
Interaction/native residual L_2 ratio	1.037
Interaction/mean-temporal L_2 ratio	1.868
Temporal-main projection	0.084
Content/order-main projection	0.916
Sequence/native distance ratio	1.100
Sequence-vector cosine	0.963

C.2 Temporal-Position and Sequence-Order Factorial Analysis

We examined temporal-position and serialized sequence-order interventions in a controlled 2×2 factorial design. The analysis contains 90 dependent episode-pair comparisons from the fixed 30-episode cohort.

The factorial intervention distinguishes the contribution of temporal representation, the contribution of serialized content/order, and their interaction. The observed action displacements were heterogeneous across episodes and were not adequately characterized by a uniform additive effect.

The descriptive medians are summarized in Table 2.

The projection measurements characterize directional alignment in action space. In particular, the median content/order-main projection of 0.916 must not be interpreted as 91.6% of variance explained.

C.3 Sequence-Order Isolation

A separate sequence-order intervention changed serialized historical order while maintaining the same semantic evidence and a common temporal assignment.

Across the 90 dependent episode-pair comparisons, the median sequence/native action-distance ratio was approximately 1.100, and the median sequence-vector cosine was 0.963. The median sequence projection was 0.916.

These measurements establish sequence-order sensitivity at the initial action-output level. They do not isolate the precise downstream operation responsible for that sensitivity. The subsequent U/V/W experiment therefore directly intervenes on serialized memory-key positions while controlling the represented-memory permutation.

D Original U/V/W Diagnostic and Numerical Controls

D.1 Original Diagnostic

The initial U/V/W experiment evaluated three selected diagnostic units, one from each task family.

Condition U used the original represented-memory arrangement and its corresponding serialized key positions. Condition V permuted the represented memory while assigning new serialized key positions. Condition W used the same represented-memory permutation as V but transported each memory unit's original key-position identity.

For FrameSamp, the permutation operated on complete 16-token retained-frame groups, preserving the association among the represented memory tokens belonging to each selected frame.

Under the original ordinary execution procedure, the strict U/W recovery criterion passed in only one of the three selected units. This original $1/3$ result is preserved as the historical diagnostic.

D.2 Post-Failure Numerical Investigation

Subsequent same-process diagnostics investigated the two units that had failed strict recovery.

The diagnostics identified sensitivity to floating-point reduction order. Even where the intended content-position association is equivalent, a change in the order of floating-point operations can produce small numerical differences (Goldberg, 1991).

After matching the reduction order, canonical W recovered the reference output bitwise in all three selected diagnostic units. This $3/3$ post-failure diagnostic does not replace or retroactively alter the original $1/3$ result.

The two observations answer different questions: the original experiment tested strict recovery under ordinary execution, whereas the subsequent diagnostic tested whether numerical execution order explained the residual exact-equality failures.

D.3 Canonical U/V/W Control

The canonical intervention evaluates three conditions under matched floating-point reduction order:

- U_c**:** Original represented memory with its original serialized key-position assignment.
- V_c**:** Permuted represented memory with key positions assigned according to the new serialized slots.
- W_c**:** The same permuted represented memory as Vc, with each memory unit's original key-position identity transported alongside it.

The corresponding attention mask is transported consistently with the represented memory. The current query observation, task instruction, checkpoint, and memory content are held fixed across conditions.

Under this control, the scientific contrast between Vc and Wc is serialized key-position assignment, not a change in the represented-memory tokens. The prospective experiment described in Appendix E uses this canonical control from the outset.

E Full Stage-7 Results

E.1 Frozen Cohort and Primary Criterion

The prospective FrameSamp experiment expands the canonical intervention from the three selected diagnostic units to the remaining 27 primary episodes of the original fixed cohort.

Three episode-0 units are retained as engineering controls and excluded from the primary scientific sample. Each primary episode is evaluated in three fresh process blocks, P1, P2, and P3.

For each episode and process block, the frozen joint criterion requires both of the following:

1. U_c and W_c must be bitwise identical in dtype, shape, and raw 20×8 action output.
2. The U_c–V_c displacement on the analyzed 16×8 action chunk must exceed at least one of the predeclared legacy numerical nuisance bounds.

The frozen FrameSamp displacement criterion is

$$\max |U_c - V_c| > 0.0019527353515624846 \quad \text{or} \quad \mathrm{MAE}(U_c, V_c) > 0.00013632749964172784.$$

where both displacement statistics are evaluated on the initial 16×8 action chunk.

The exact-recovery requirement is evaluated on the full raw 20×8 output and is therefore distinct from the displacement threshold.

Table 3: Prospective FrameSamp results. The scientific denominator is 27 fixed primary episodes; process-level checks assess reproducibility on those episodes.

Criterion	Result
Primary episodes passing joint gate	$27/27$
U_c–W_c raw-output exact recovery	$81/81$
U_c–V_c displacement above nuisance bound	$81/81$
Fresh process blocks	3
Engineering controls excluded from cohort	3

Table 4: Executed downstream bridge estimates. Overall and control-relative intervals use the prespecified task-stratified episode bootstrap; task-specific intervals use within-task episode resampling.

Contrast	Estimate	95% CI
Overall FAR alignment	$+0.002948$	$[+0.001999, +0.004049]$
FAR minus common shuffle	$+0.001300$	$[+0.000663, +0.001917]$
FAR minus mismatched shuffle	$+0.002089$	$[+0.001377, +0.002835]$
VideoUnmask	$+0.006965$	$[+0.006282, +0.007729]$
InsertPeg	$+0.002913$	$[+0.000297, +0.005976]$
VideoPlaceOrder	-0.001035	$[-0.001872, -0.000302]$

E.2 Prospective Outcome

All 27 fixed primary episodes satisfy the joint criterion across the three fresh process blocks.

In all 81 episode–process checks, W_c recovers U_c bitwise at the raw action-output level. In the corresponding 81 checks, V_c remains displaced from U_c beyond the predeclared nuisance criterion.

The result is reported as a census of the fixed 27-episode cohort. The repeated process blocks do not create 81 independent scientific episodes. Accordingly, no population-level significance test is attached to the exact mechanistic gate.

E.3 Interpretation and Numerical Scope

The prospective experiment demonstrates that transporting original serialized key positions is sufficient to recover the reference action under the canonical matched-reduction-order control in this fixed cohort.

The corresponding Vc displacement establishes action-level dependence on serialized key-position assignment in the tested downstream interface.

The experiment does not establish general invariance under arbitrary memory transformations or guarantee that the displaced actions reduce closed-loop task success. Those are distinct scientific questions.

F Executed Downstream Bridge Details

The bridge uses the 27 primary retention-matched episodes and evaluates each at inference seeds 42, 7, and 2026. For a matched layout, the selected frame identities and upstream represented memory are held fixed. Each semantic frame identity is mapped to its rank in the FAR anchor and expanded to 16 consecutive memory-token coordinates. Duplicate or missing identities are fatal protocol errors. A post-RoPE inverse-identity gather keeps reduction order matched across the canonical comparisons.

The inferential unit is the episode. The three seeds are repeated inference evaluations, not 81 independent tasks. Confidence intervals use a task-stratified episode bootstrap with 20,000 replicates. Positive estimates indicate a reduction in action divergence relative to the stated reference.

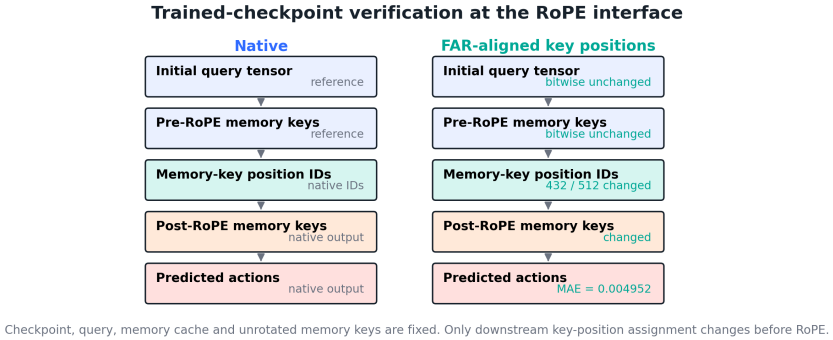

Figure 10: Trained-checkpoint verification at the RoPE interface. Initial queries and pre-RoPE memory keys are unchanged; only downstream memory-key position IDs are reassigned before rotation, after which rotated keys and predicted actions differ.

The runtime interface probe uses an identical matched-history input across native and FAR-aligned conditions. Initial queries, query-position IDs, memory cache, and unrotated memory keys are bitwise unchanged. The bridge changes 432 of 512 memory-key position IDs immediately before RoPE, after which rotated keys and predicted actions differ (action MAE 0.004952). The probe executes no simulator steps and is therefore mechanistic evidence, not a behavioral evaluation.

G TokenDrop Cross-Representation Results

G.1 Representation and Intervention

The TokenDrop-Modul experiment tests the canonical U/V/W intervention under a different upstream perceptual-memory representation.

FrameSamp permutes complete 16-token retained-frame groups. TokenDrop instead permutes individual selected sparse memory tokens. Both variants retain the tested downstream memory-as-modulation interface.

The TokenDrop intervention preserves the same three-condition causal structure: U_c is the reference condition, V_c assigns new serialized key positions to the permuted represented memory, and W_c transports the original positions with that same permutation.

The current query, task instruction, and represented memory are fixed across the corresponding U/V/W comparisons.

G.2 Evaluation Cohort and Engineering Controls

TokenDrop uses the same 27 fixed primary episodes as the prospective FrameSamp experiment and three fresh process blocks per episode.

The three episode-0 engineering controls are excluded from the scientific sample. An initial engineering validation exposed a configuration issue. The subsequent protocol amendment and successful validation preceded the primary fixed-cohort evaluation; the unsuccessful initial validation is not counted as scientific evidence.

G.3 Replication Outcome

All 27 primary TokenDrop episodes satisfy the canonical joint criterion across the three fresh process blocks.

Under both upstream representations, transporting original serialized positions recovers the reference action under canonical numerical control, whereas assigning new serialized positions preserves action displacement.

Table 5: Fixed-cohort positional-transport outcomes under the two upstream perceptual-memory representations.

Criterion	FrameSamp	TokenDrop
Primary episodes passing	$27/27$	$27/27$
U_c–W_c exact checks	$81/81$	$81/81$
U_c–V_c displaced checks	$81/81$	$81/81$
Fresh process blocks	3	3

Table 6: Prospective closed-loop success counts. Each condition contains 30 paired rollout units.

Condition	Successful rollouts
FAR_NATIVE	$8/30$
CLEAN	$7/30$
FAR_TPOS_RECENT	$5/30$
RECENT_NATIVE	$4/30$

The replication therefore establishes that the observed effect is not restricted to FrameSamp's particular retained-frame construction. Because both variants share the same downstream modulation interface, the result does not establish generalization across independent architecture families.

H Closed-Loop Evaluation Details

The main text reports the historical upstream success-rate visualization in Figure 8 and the downstream development cohort in Figure 9. The appendix retains the historical paired tests and records the downstream intervention separately because the two studies manipulate different positional interfaces.

H.1 Prospective Paired Rollout Design

A separate prospective experiment evaluates the behavioral consequences of native temporal-history placement and an upstream temporal-position correction.

Thirty paired rollout units are evaluated under four conditions:

- CLEAN: The ordinary non-intervened baseline for the selected rollout setting.
- FAR_NATIVE: The native FAR history condition.
- RECENT_NATIVE: The native RECENT history condition.
- FAR_TPOS_RECENT: The FAR history is retained while the selected historical frames receive the temporal embedding positions corresponding to RECENT.

The four conditions yield 120 completed rollouts, with 30 rollouts per condition.

This experiment intervenes on upstream per-frame temporal representation. It does not evaluate downstream U/V/W serialized-key-position transport under closed-loop feedback.

H.2 Success Counts

The two pre-specified improvement comparisons use FAR_NATIVE as their paired reference. Exact McNemar tests assess the paired binary outcomes, with Holm correction applied across the two planned comparisons.

Here, pp denotes percentage points. The planned comparisons do not establish an improvement over the FAR native condition.

Table 7: Planned paired closed-loop comparisons. Differences are relative to FAR_NATIVE. Neither comparison confirmed the hypothesized improvement.

Comparison	Difference	Exact p	Holm p
RECENT vs FAR	-13.3 pp	.125	.250
TPOS correction vs FAR	-10.0 pp	.375	.375

Table 8: Downstream closed-loop development results. The intervention was active on every applicable age-arm policy call, but no paired official outcome changed.

Quantity	Result
Paired task episodes	15
Official rollouts	45
Native successes	$7/15$
Identity-control successes	$7/15$
Age-position successes	$7/15$
Age-arm calls with intervention active	$293/293$
Paired rescues vs. native / identity	$0/0$
Paired regressions vs. native / identity	$0/0$

H.3 Task-Family Heterogeneity

Closed-loop success is heterogeneous across the evaluated task families. In particular, InsertPeg exhibits a floor effect, with no successful rollouts among its 10 episodes under any of the four conditions.

The other task families exhibit different success patterns. The task-family observations are descriptive and are not treated as independent confirmatory subgroup tests.

The lack of confirmed improvement in this 120-rollout experiment does not establish the absence of every possible behavioral effect. Nor does it contradict the independently evaluated downstream U/V/W action-level result.

H.4 Downstream Age-Position Development Cohort

The downstream study contains 15 paired episodes across VideoUnmask, InsertPeg, and VideoPlaceOrder. Each episode is reset and evaluated under native, identity-control, and age-position arms. The evaluator executes 16 actions per policy chunk and buffers only subsequently observed frames. The online mapping uses acquisition timestamps already available to the policy and receives no privileged relevance or success signal.

Because the cohort combines pilot and development episodes and contains no outcome disagreements, it supports an operational negative result rather than a held-out efficacy or equivalence claim.

I Implementation Details and Reproducibility

I.1 Fixed Experimental Configuration

The principal FrameSamp experiments use the RoboMME FrameSamp-Modul checkpoint

`Yinpei/perceptual-framesamp-modul/79999`.

Unless otherwise specified, inference uses seed 42. The FrameSamp memory configuration provides 512 perceptual-memory tokens, corresponding to 32 retained historical frames with 16 tokens per frame.

The original fixed evaluation cohort includes the following episode identifiers in each of the three task families:

$$\{0, 5, 11, 16, 22, 27, 33, 38, 44, 49\}.$$

The three task families are VideoUnmask, InsertPeg, and VideoPlaceOrder.

The prospective mechanistic cohort excludes episode 0 from each family, yielding 27 primary scientific episodes. The excluded units serve as engineering controls.

I.2 Experimental Provenance

The experimental workflow distinguishes frozen scientific decisions from subsequent engineering corrections. The original Stage-4D strict-recovery failure is retained rather than overwritten. Subsequent numerical diagnostics motivate the separately frozen canonical intervention.

During preparation of the prospective FrameSamp evaluation, historical frame-pixel comparisons identified localized reconstruction differences. A versioned provenance amendment therefore distinguished current-run frozen rendering hashes from historical reference hashes, while preserving the historical discrepancies as diagnostic records.

The provenance and integrity checks are separate from the action-level joint criterion. An infrastructure or integrity failure is not silently counted as a scientific pass or scientific negative.

The TokenDrop experiment likewise preserves the unsuccessful initial engineering validation and distinguishes it from the primary evaluation performed under the amended protocol.

The executed bridge and downstream rollout study use versioned manifests, immutable per-run receipts, source and checkpoint hashes, and separate infrastructure-error records. Inference seeds 42, 7, and 2026 are recorded as repeated evaluations of the same 27 bridge episodes. The 45-rollout cohort is explicitly labeled development data.

I.3 Numerical Reproducibility

The canonical U/V/W experiment controls the floating-point reduction order used for the reference and intervention conditions.

The equality criterion compares the raw 20×8 action output, including its dtype, shape, and exact values. The separate action-displacement criterion uses the initial 16×8 chunk.

All repeated-process results are interpreted as reproducibility checks on fixed scientific episodes, not additional independent samples.

I.4 Artifact Requirements for Independent Verification

Independent verification of the reported experiments requires the executed protocol and runner versions, checkpoint identification, frozen cohort and intervention manifests, original per-episode action arrays, engineering control outcomes, and process-level execution logs.

For the retention-matched experiment, the original selected-frame identities and natural and matched action arrays are necessary to reconstruct the paired statistical tests. Derived summary tables alone do not independently establish the underlying action comparisons.

For the prospective mechanistic experiments, the raw Uc, Vc, and Wc outputs, the frozen numerical nuisance criteria, permutation manifests, and per-process integrity checks are required to reconstruct the exact-recovery and action-displacement gates.

For the executed bridge, verification additionally requires the three-seed per-episode action bundles, selected-frame identities, FAR-aligned and shuffle position maps, canonical reduction-order traces, and trained-checkpoint RoPE captures.

For the historical closed-loop evaluation, verification requires episode-aligned binary outcomes under all four conditions, rather than marginal success counts alone. The paired outcomes are necessary to reproduce the exact McNemar tests. The downstream development study additionally requires native/identity/age rollout receipts, per-call intervention logs, and official task-outcome records for all 15 triplets.

The distinction between raw experimental artifacts and derived analysis outputs is maintained when interpreting all reported results.