MEMORY-CARRIER CONTRAST SET: ISOLATING TEMPORAL CARRY IN FROZEN LANGUAGE-MODEL CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon control with a frozen language model poses an attribution problem: a recurrent adapter may win because it transports evidence across decisions, because it changes the policy's current-feature access, or because its training contract differs from the retained-context baseline. The Memory-Carrier Contrast Set (MCS) turns these alternatives into explicit interventions on carry and access while retaining backbone features, supervision, head geometry, optimizer, and interaction budget. We instantiate MCS with Layered State Bottleneck (LSB), which mixes stop-gradient features from three transformer depths into a 128-dimensional GRU state that alone drives the decision heads. Across prespecified persistent-history tasks in MiniGrid-S13, TextWorld, and ALFWorld, the matched carry-by-access factorial assigns 4.1 success points (95% CI [2.9, 5.3]) to temporal carry; without bypass, the task-wise effects are 4.7–5.5 points, while the S9/S11 effects remain inside a two-point equivalence margin. Under a matched LLaMA-2-7B contract, LSB also exceeds a native FLC-4096 controller by 5.4 points ([3.9, 6.9]); the measured uncached full-prefix implementation processes 17.6 times as many retained tokens per decision and has 2.8 times the p95 latency. A final-layer GRU reproduces the concentration on persistent-history tasks, and task-only LSB retains a 2.6-point advantage over projected bypass while remaining within 0.3 points of the DTQN-style encoder. MCS therefore identifies temporal transport as the active ingredient in this frozen-controller regime and places bounded recurrence on a favorable success–resource operating point.

1 INTRODUCTION

A recurrent state can outperform a frozen long-context policy while leaving the scientific question unanswered. Recurrence transports past evidence, but recurrent systems also alter trainable geometry, restrict or expose current backbone features, and often use different labels or interaction budgets. A configuration-level return gap bundles these changes together. For long episodes, that ambiguity obscures both why the policy improves and whether bounded state is preferable to repeatedly processing a retained prefix.

Memory evaluation for RL has recently formalized correlation horizon and the context-memory border to distinguish short- from long-term demands (Cherepanov et al., 2026a). That framework determines whether a task and context window expose a memory requirement. It does not identify which information path produces a gain when recurrent carry, direct feature access, supervision, and capacity change together. Existing recurrent and transformer policies for POMDPs likewise establish competitive model families without a carry-by-access intervention inside a frozen language-model controller (Ni et al., 2022; Lu et al., 2024).

The Memory-Carrier Contrast Set (MCS) targets this missing causal contrast. Its core experiment crosses carried state with a projected current-feature bypass, then resets recurrence while holding frozen features, losses, augmented head geometry, optimizer, and budget fixed. Separate coordinates compare complete configurations under native context and task-only supervision, with backbone and budget labels preserved. Layered State Bottleneck (LSB) provides the bounded carrier: stop-gradient reads from three transformer depths are mixed into a 128-dimensional GRU, and its state is

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 exclusive input to policy, value, and recovery heads. MCS thereby turns "memory helps" into estimands for temporal transport, access, configuration-level context, and measured resource cost.

The empirical pattern is sharp. In the matched GPT-2 factorial, carry contributes 4.1 points (95% CI [2.9, 5.3]) over MiniGrid-S13, TextWorld, and ALFWorld; disabling bypass yields positive task-wise effects of 4.7–5.5 points, while S9/S11 remain inside the prespecified equivalence margin. With LLaMA-2-7B, LSB exceeds native FLC-4096 by 5.4 persistent-task points and reduces measured uncached full-prefix p95 latency by 65%. The concentration persists with a final-layer GRU, and a task-only LSB remains competitive with trainable history encoders. These findings support three contributions:

- **Memory-carrier identification.** MCS defines a carry-by-access factorial, contract residuals, and an interval-contained equivalence rule that separate temporal transport from current-feature access and configuration-level comparisons.
- **Persistent-history mechanism evidence.** The controlled factorial isolates a practically meaningful carry effect on every persistent-history task, while task-wise intervals and a final-layer recurrent readout establish the task-group concentration.
- **A success–resource operating map.** Native-context, task-only, and trainable-history comparisons jointly expose success, retained tokens, latency, inference memory, parameter, interaction, and training-compute costs without averaging incompatible contracts.

2 Related Work

Recurrent model-free RL and transformer policies provide strong but differently biased POMDP baselines (Hausknecht & Stone, 2015; Parisotto et al., 2020; Ni et al., 2022; Lu et al., 2024). Transformer-XL, recurrent memory transformers, structured state-space models, and memory traces extend temporal range or compress a trained sequence model's history (Dai et al., 2019; Bulatov et al., 2022; Gu et al., 2022; Eberhard et al., 2025). Most directly, Cherepanov et al. (2026a) use correlation horizons and a context-memory border to configure tests of short- and long-term memory. MCS begins after that demand has been designated: it estimates the effect of carrying state while holding the frozen representation and training contract fixed. The crossed bypass further distinguishes temporal transport from direct current-feature access, an ambiguity that horizon classification and model-family comparisons do not resolve.

Parameter-efficient adaptation changes where trainable capacity enters a frozen model. Adapters, LoRA, prefix tuning, and prompt tuning use residual, low-rank, or input-side trainable state (Houlsby et al., 2019; Hu et al., 2022; Li & Liang, 2021; Lester et al., 2021; He et al., 2022). READ recurrently processes intermediate representations across transformer depth in a side network, rather than carrying state across environment decisions (Nguyen et al., 2023). LSB instead mixes multiple frozen depths into a temporally recurrent state that is the exclusive input to its decision heads. The projected-bypass and final-layer-GRU controls separately test restricted policy access and recurrent feature source.

DTQN and RATE train history encoders directly (Esslinger et al., 2022; Cherepanov et al., 2026b); the final RATE architecture combines segment memory with a Memory Retention Valve in offline RL. Our external controls adapt the published DTQN and segment-recurrent carrier families to the paper's common online PPO interface, so we label them DTQN-style and RATE-style rather than reproductions of the original learning algorithms. Latent-state and world-model methods instead learn predictive state or dynamics end to end (Kaelbling et al., 1998; Deisenroth & Rasmussen, 2011; Igl et al., 2018; Hafner et al., 2019; Zintgraf et al., 2020; Hafner et al., 2020; Micheli et al., 2023). MCS measures temporal transport with a frozen representation before locating that controlled effect among complete retained-context and trainable-history configurations.

Table 1: Closest memory families and the intervention added by MCS. "Access crossed" means that direct current-feature access is changed independently of recurrent carry.

Family	Backbone and carrier	Supervision and access	Primary comparison
Adapters, LoRA, prefix	Frozen/light; residual module or tokens	Task targets; backbone path usually available	Adaptation interface
READ	Frozen; depth-recurrent side tuner	Task targets; backbone path available	Recurrent PEFT across layers
Temporal GRU heads	Frozen; final-layer temporal state	Task or auxiliary targets; state controls policy	Recurrent control across decisions
Memory tokens and RMT	Frozen or retrained; token state	Task targets; access varies	Long-history quality
DTQN and RATE	Retrained history encoder	Task targets; encoder controls policy	End-to-end control return
Memory traces	Trained compact trace	Task targets; trace controls policy	Window versus compressed state
Memory evaluation	Agent-dependent	Horizon/context-border protocol	Memory-demand classification
LSB + MCS	Frozen multi-depth GRU	Supervision matched/removed; access crossed	Carry effect, FLC-4096 residual, resources

3 METHOD: MEMORY-CARRIER CONTRAST SET

Every MCS coordinate is indexed by an evaluation contract C and a prespecified task subset E. C_G denotes GPT-2 Medium with auxiliary supervision, 5M steps, and matched frozen feature geometry; C_L denotes LLaMA-2-7B with auxiliary supervision, 5M steps, and its native 4096-token window; C_T denotes the task-only 5M-step comparison; and C_F denotes the set of explicitly labeled backbone–supervision–budget operating points used for the resource frontier.

3.1 BOUNDED RECURRENT INSTANTIATION

We use standard POMDP notation: latent state s_t generates observation o_t, and the policy acts from causal history. Let $F = (F^1, \dots, F^L)$ be a frozen transformer and $\mathcal{L}_{\text{ins}}$ the readout layers. At each selected layer, LSB projects the frozen feature $\mathbf{h}_t^l$ into a d-dimensional summary:

$$\mathbf{x}_t^l = W_l^{\text{proj}}\, \text{sg}\big(\mathbf{h}_t^l\big)\,, \qquad \mathbf{x}_t = W_{\text{mix}}[\mathbf{x}_t^l]_{l\in\mathcal{L}_{\text{ins}}}. \tag{1}$$

The stop-gradient operator prevents updates to the backbone. A GRU transports the mixed evidence through time:

$$\begin{aligned} \mathbf{z}_t &= \sigma(W_z[\mathbf{x}_t; \mathbf{b}_{t-1}] + \mathbf{c}_z), & \mathbf{r}_t &= \sigma(W_r[\mathbf{x}_t; \mathbf{b}_{t-1}] + \mathbf{c}_r), \\ \widetilde{\mathbf{b}}_t &= \tanh(W_b[\mathbf{x}_t; \mathbf{r}_t \odot \mathbf{b}_{t-1}] + \mathbf{c}_b), & \mathbf{b}_t &= (1-\mathbf{z}_t)\odot \mathbf{b}_{t-1} + \mathbf{z}_t \odot \widetilde{\mathbf{b}}_t. \end{aligned} \tag{2}$$

Policy $\pi_\psi(a \mid \mathbf{b}_t)$, value $V_\xi(\mathbf{b}_t)$, and recovery $g_\phi(\mathbf{b}_t)$ heads receive only $\mathbf{b}_t$; no head accesses raw backbone features. This information-flow contract distinguishes LSB from final-layer recurrence and input-side memory while retaining the same frozen feature source across the factorial controls.

The two bypass controls test distinct access paths. The supervision-matched P/V bypass forms $\mathbf{q}_t = Q\,\text{sg}(\mathbf{h}_t^L) \in \mathbb{R}^d$ from the final frozen token and supplies $[\mathbf{b}_t; \mathbf{q}_t]$ to policy and value while recovery still receives only $\mathbf{b}_t$. The factorial's projected bypass instead supplies $[\mathbf{b}_t; \mathbf{x}_t]$, exposing the same multi-depth mixture that enters the GRU. Every factorial row uses the same augmented head geometry; bypass-off supplies a zero vector of the projected-bypass width, including the displayed carry-on LSB anchor. "Reset" replaces $\mathbf{b}_{t-1}$ by zero at every training and evaluation step in Equation 2; all other computation is unchanged.

3.2 TRAINING CONTRACT

The simulator-supervised contract combines PPO (Schulman et al., 2017) with state recovery, contrastive separation, and calibration:

$$\mathcal{L} = \mathcal{L}_{\text{PPO}} + \alpha\mathcal{L}_{\text{state}} + \beta\mathcal{L}_{\text{contrast}} + \gamma\mathcal{L}_{\text{calib}}. \tag{3}$$

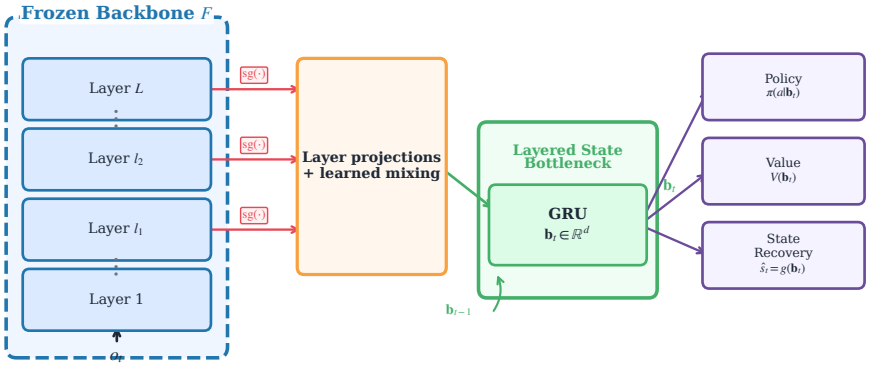


Figure 1: LSB instantiates the bounded-state carrier: stop-gradient multi-depth features update a recurrent state that alone controls the heads. The factorial independently intervenes on temporal carry and projected bypass permission.

State recovery regresses action-relevant latent factors, InfoNCE groups equal latent states across distinct histories while separating observation-aliased states (van den Oord et al., 2018), and calibration maximizes recovery entropy when a complete causal prefix remains observationally ambiguous. Table 7 gives broad reference configurations, with LSB using the auxiliary package and the local references using PPO alone. All attribution claims instead use controlled rows that receive identical auxiliary targets and change only the named access, recurrence, or feature-source intervention. Simulator state is never available to the policy at evaluation time. Appendix A gives target definitions and pair construction.

3.3 Contrast set and decision rule

For method m and task subset E, let $S^C(m, E)$ denote aggregate mean success under contract C. The retained-context residual and matched contract-residual vector are

$$R^{C_L}_{\text{ctx}}(E) = S^{C_L}(LSB, E) - S^{C_L}(\text{FLC}_{4096}, E),$$
$$\mathbf{R}^{C_G}_{\text{ctr}}(E) = \left\langle S^{C_G}(LSB, E) - S^{C_G}(\text{PVBypass}, E), S^{C_G}(LSB, E) - S^{C_G}(\text{ProjBypass}, E) \right\rangle. \tag{4}$$

The MCS profile for E is

$$\mathcal{P}(E) = \left\langle \mathbf{R}^{C_G}_{\text{ctr}}(E), \Delta^{C_G}_{\text{carry}}(E), R^{C_L}_{\text{ctx}}(E), \mathcal{F}^{C_F}(E) \right\rangle, \tag{5}$$

where $\Delta^{C_G}_{\text{carry}}$ is the factorial carry effect and $\mathcal{F}^{C_F}$ is the nondominated set over success, supervision availability, retained tokens, trainable parameters, interaction budget, p95 latency, inference memory, and GPU-hours; every point retains its contract label. $R^{C_L}_{\text{ctx}}$ compares complete carrier configurations under C_L, while the matched factorial isolates temporal transport inside C_G. Before training, we fix two subsets from environment structure: the two smallest single-junction MiniGrid variants S9/S11 form E_{short}, while MiniGrid-S13 and the cross-transition inventory, quest, and object-state demands of TextWorld and ALFWorld form $E_{\text{persistent}}$. No observed return enters this partition.

The prespecified two-point practical margin is interpreted with interval logic. We call two configurations practically equivalent only when the entire 95% interval for their difference lies inside $[-2, 2]$ success points, and a practically meaningful recurrent advantage requires the interval to lie above two points. An interval above zero but crossing two supplies directional, rather than practical-threshold, evidence; remaining estimates are reported without a category. Two points correspond to 20 additional solves per 1,000. This rule makes the S9/S11 FLC comparison unresolved because its interval extends to 2.9, while the persistent-task FLC and factorial carry effects clear the practical threshold. For resource points p and q, p dominates q only if it uses no stronger supervision, attains no lower success, and uses no more retained tokens, trainable parameters, interaction steps, p95 latency, inference memory, or GPU-hours, with a strict improvement on at least one coordinate. The nondominated points form $\mathcal{F}^{C_F}$.

4 EXPERIMENTS: IS CARRY THE ACTIVE INGREDIENT?

4.1 TASKS, MODELS, AND EVALUATION PROTOCOL

The empirical sequence follows the identification claim. Q1 estimates temporal carry after crossing it with current-feature access. Q2 tests whether the resulting task-group pattern accompanies a configuration-level advantage over a native 4096-token controller and measures the associated resource point. Q3 changes recurrent readout and supervision to test how broadly the persistent-history concentration survives.

MiniGrid Memory (Chevalier-Boisvert et al., 2023) provides S9, S11, and S13 horizons of approximately 9–40 steps. TextWorld (Côté et al., 2019) uses five procedurally generated games that require persistent object and quest state. ALFWorld (Shridhar et al., 2021) covers six task types and 134 environments with latent object, receptacle, appliance, and progress state. All methods consume the same benchmark-native serialized observation/action stream and use native success criteria.

Each transition is serialized oldest-to-newest as task description, current observation, and previous action, separated by fixed field delimiters and encoded with the backbone's native tokenizer. LSB encodes only the current transition block and reads the final non-padding token at each selected layer; FLC concatenates transition blocks and left-truncates to the most recent 2048 or 4096 tokens. Recurrent states are zeroed at episode boundaries and detached at each 256-step rollout boundary. Policies score the benchmark-provided admissible action set, with identical action interfaces and native episode caps. MiniGrid has no dataset-level train/evaluation split: after training, each frozen policy is evaluated on 200 fresh resets from the environment's procedural reset distribution per seed. TextWorld uses the five held-out games listed in Appendix A, with 40 episodes per game and seed; ALFWorld uses its fixed 134-environment out-of-distribution evaluation split.

GPT-2 Medium has 355M parameters, width 1024, and 24 layers (Radford et al., 2019); LSB reads layers $\{8, 16, 23\}$ with $d = 128$. LLaMA-2-7B has width 4096 and 32 layers (Touvron et al., 2023); reads are $\{8, 16, 31\}$. Local comparators are Frozen Long-Context (FLC), 16 learned memory tokens per step, selective fine-tuning of the top three layers, and two bypass controls. The GPT-2 last-layer GRU comparator has width 128; the LLaMA cross-workflow replication uses width 256. DTQN-style and RATE-style retrain their history encoders end to end.

FLC applies the same policy/value heads to the final frozen token and carries no state outside its retained prefix. The GRU-head control recurrently updates from that final token; memory-token control prepends its learned tokens to each current transition; and selective fine-tuning updates only the top three transformer layers. The trainable-history controls isolate encoder choice under the common online interface: DTQN-style uses a causal 2048-token transformer, while RATE-style passes 16 memory tokens between 512-token segments. Both train the encoder and action-query head end to end with task-only PPO on the common stream. These adaptations retain the published carrier geometry while replacing DTQN's Q-learning and RATE's offline return-conditioned MRV algorithm with one online objective (Esslinger et al., 2022; Cherepanov et al., 2026b). Appendix A gives the complete adaptations and auxiliary-target construction.

PPO uses Adam at 3×10^{-4} with $\epsilon = 10^{-5}$, 128 parallel environments, 256-step rollouts, four epochs, and 5M environment steps unless a larger comparator budget is displayed. The auxiliary weights are $\alpha = 0.5$, $\beta = 0.1$, and $\gamma = 0.05$; all task-only rows set them to zero. Independent training seeds are $\{11, 23, 47, 89, 131\}$, and 200 evaluation episodes per seed and task yield 1,000 episodes per method–task cell. Task cells report seed means and standard deviations; aggregate columns report means unless an aggregate deviation is shown. Pairwise intervals resample training seeds and episodes hierarchically for shared-seed designs and use unpaired hierarchical resampling for separately trained history encoders; 10,000 draws form percentile 95% intervals. Because repeats are not additionally clustered by TextWorld game or ALFWorld environment, these intervals condition on the five named games and fixed 134-environment split rather than extrapolating to unseen instance populations. Contrasts use unrounded task and seed values. Training uses eight NVIDIA A100 80GB GPUs, CUDA 12.4, PyTorch 2.2, and BF16.

Table 2: Factorial effects under C_G. All four cells share head geometry, frozen features, auxiliary losses, optimizer, stream, and 5M-step budget.

Persistent-history effect	Estimate	95% CI	Empirical role
Recurrent carry	4.1	[2.9, 5.3]	Temporal transport
Projected bypass	-1.6	[-2.6, -0.7]	Access path
Carry × bypass	1.8	[0.3, 3.4]	Carry–access coupling

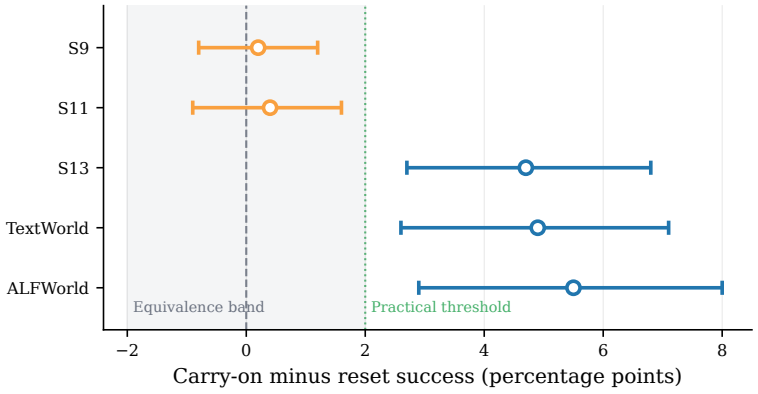


Figure 2: Task-wise carry effects with bypass disabled under C_G. Intervals lie inside the two-point equivalence band on S9/S11 and above the practical threshold on all three persistent-history tasks.

4.2 Q1: Isolating temporal transport

The four factorial cells make the intervention visible before any model-family comparison. Persistent-history success is 74.4 with carry and no bypass, 71.8 with both paths, 69.3 after resetting carry without bypass, and 68.6 after resetting carry with bypass (Appendix Table 9). For success Y_{cb} with carry c and bypass b, the predefined effects are

$$\begin{aligned} \Delta_{\text{carry}} &= \tfrac{1}{2}[(Y_{10} - Y_{00}) + (Y_{11} - Y_{01})], \\ \Delta_{\text{bypass}} &= \tfrac{1}{2}[(Y_{11} - Y_{10}) + (Y_{01} - Y_{00})], \\ \Delta_{\text{int}} &= (Y_{10} - Y_{00}) - (Y_{11} - Y_{01}). \end{aligned} \tag{6}$$

Figure 2 shows why the 4.1-point main effect is consequential rather than an aggregate artifact. Carry changes S9/S11 by only 0.2 and 0.4 points, with both intervals contained inside the equivalence margin, but improves S13, TextWorld, and ALFWorld by 4.7–5.5 points without bypass. The positive interaction shows that restricting direct current-feature access increases the value of carried evidence.

Two additional matched contrasts locate this effect within frozen-backbone design choices. The persistent-history contract-residual vector is $\langle 3.1, 2.5 \rangle$ points against P/V and projected bypass, with intervals [1.8, 4.5] and [1.3, 3.8]. A matched last-layer GRU reaches 76.1 five-task average success, compared with LSB at 80.9 (Appendix Table 8). The width sweep peaks at $d = 128$ rather than increasing monotonically (Appendix H).

4.3 Q2: Comparison with a native 4096-token controller

The persistent-history residual is 5.4 points (95% CI [3.9, 6.9]), with all five seed-level subset averages favoring LSB. This equals 46–61 additional solves per 1,000 on S13, TextWorld, and ALFWorld. The S9/S11 residual is smaller at 1.4 points ([-0.1, 2.9]); because the upper endpoint exceeds two points, MCS records this comparison as unresolved rather than as evidence of equivalence. Complementary GPT-2 context endpoints appear in Appendix Table 12.

Table 3: The recurrent-state configuration exceeds native FLC-4096 by 4.6–6.1 points on the pre-specified persistent-history tasks. The S9/S11 aggregate interval crosses the two-point equivalence boundary.

Task	FLC-4096	LSB	$R^{C_L}_{\mathrm{ctx}}$
S9	94.1 ± 1.2	95.0 ± 1.0	0.9
S11	87.3 ± 1.6	89.2 ± 1.4	1.9
S13	79.0 ± 1.9	83.6 ± 1.7	4.6
TextWorld	72.8 ± 2.1	78.2 ± 1.9	5.4
ALFWorld	68.0 ± 2.3	74.1 ± 2.2	6.1

Table 4: Contract-labeled operating points on eight A100 80GB GPUs. Runtime is measured at batch one; FLC re-encodes its retained prefix without a cross-decision KV cache.

Method	Supervision	Steps (M)	Persist.	Tokens/dec.	p95 (ms)
LLaMA LSB	Auxiliary	5	78.6	186	46.7
LLaMA FLC-4096	Auxiliary	5	73.3	3276	132.4
LLaMA proj. bypass	Auxiliary	5	76.7	186	50.8
DTQN-style	Task-only	10.0	72.0	768	27.9
RATE-style	Task-only	12.0	74.6	512	35.8

Method	Trainable (M)	Total model	Infer. mem. (GB)	GPU-hours
LLaMA LSB	1.94	$\sim$7B	17.3	612
LLaMA FLC-4096	0.58	$\sim$7B	24.7	1328
LLaMA proj. bypass	2.47	$\sim$7B	17.8	657
DTQN-style	47.6	47.6M	8.4	1184
RATE-style	72.8	72.8M	10.9	1646

4.4 Measured performance and resource cost

Against the measured uncached full-prefix implementation, LSB combines a 5.4-point persistent-history advantage with 17.6 times fewer retained tokens, 65% lower p95 latency, 30% lower inference memory, and 54% fewer GPU-hours. The DTQN-style and RATE-style encoders are faster and use less inference memory, whereas LSB trains 24.5 and 37.5 times fewer parameters with half or less of their interaction budgets and lower total GPU-hours. The useful choice is therefore workload-dependent, but bounded frozen-state adaptation occupies a distinct high-success, low-trainable-state point.

4.5 Q3: Readout and supervision robustness

The cross-workflow test replaces LSB's multi-depth recurrent readout with a 256-dimensional GRU over the final frozen token representation; the policy receives only the GRU state. Auxiliary targets, PPO budget, seeds, and evaluation episodes remain fixed for the matched row, relocating recurrent integration and its feature source to the final layer.

Moving recurrence to the final layer reduces the residual's magnitude but preserves its concentration on persistent-history tasks: 3.3 points ([2.0, 4.6]) versus 0.4 ([-0.6, 1.3]). Even without the GRU's auxiliary labels, the persistent-history residual remains 1.7 points ([0.2, 3.2]). Full task values, counts, and GPT-2 results appear in Appendix Table 16.

Removing simulator-state targets from LSB yields 67.9% persistent-task success, 2.6 points above projected bypass and within 0.3 points of the DTQN-style encoder (Appendix Table 14). At the matched 5M-step budget, the RATE-style encoder leads by 1.4 points with an interval that includes zero. The separate full-budget rows are descriptive operating points because both supervision and interaction budget differ: DTQN-style solves six more S13 episodes, whereas LSB solves 29 more TextWorld and 48 more ALFWorld episodes per 1,000; RATE-style leads by 23 and one on S13 and TextWorld, while LSB leads by 18 on ALFWorld. These reversals place all three methods on distinct success–resource operating points.

Table 5: A last-layer GRU preserves the task-group contrast on LLaMA-2-7B. Residuals use the auxiliary-trained native FLC-4096 comparator; the task-only GRU therefore gives FLC a supervision advantage.

Workflow	Short residual [CI]	Persistent residual [CI]
LSB + auxiliaries	1.4 [-0.1, 2.9]	5.4 [3.9, 6.9]
GRU + auxiliaries	0.4 [-0.6, 1.3]	3.3 [2.0, 4.6]
GRU, task-only	-0.4 [-1.5, 0.7]	1.7 [0.2, 3.2]

4.6 Synthesis across contracts

Across contracts, the same empirical organization remains visible without averaging unlike comparisons. Inside C_G, the carry effect and both bypass residuals are positive on persistent-history tasks, while S9/S11 effects are small. Inside C_L, the native FLC-4096 residual grows to 5.4 points and final-layer recurrence reproduces a 3.3-point residual. Under task-only training, LSB and DTQN-style differ by 0.3 persistent-task points ([-1.3, 1.9]). MCS thus yields one identified mechanism result and a separate map of complete operating points; Appendix Tables 14 and 15 retain the full task-level comparisons.

5 Discussion

The experiments establish a mechanism result and an operating-point result at different levels. Inside C_G, resetting recurrence while preserving features, losses, augmented head geometry, and optimizer identifies temporal carry as a 4.1-point contributor on the persistent-history group. The 1.8-point carry–bypass interaction further shows that direct current-feature access changes how effectively a policy uses carried evidence. Across complete configurations, the LLaMA residual and its final-layer-GRU replication show that the persistent-history concentration is not tied to one recurrent feature source.

The resource measurements give this success difference operational scale. Relative to measured uncached FLC-4096, LSB combines 5.4 additional persistent-task success points with 17.6 times fewer retained tokens, 65% lower p95 latency, lower inference memory, and fewer GPU-hours. DTQN-style and RATE-style attain lower decision latency and memory footprint; LSB instead trains substantially fewer parameters with smaller interaction budgets. The resulting frontier makes bounded frozen-state adaptation a concrete alternative when trainable state and repeated prefix processing are the binding costs.

The evaluated regime spans discrete-action spatial, textual, and embodied simulators, two frozen decoder-only backbones, and five training seeds. Its strongest causal statement is the auxiliary-supervised C_G carry factorial; the task-only rows establish competitive complete configurations but do not separately re-estimate a label-free carry effect. Likewise, FLC comparisons measure complete carrier configurations rather than evidence known to lie outside the 4096-token window, and the latency result applies to full-prefix re-encoding, while incremental KV caching would exchange recomputation for cache memory. These boundaries preserve a clear decision consequence: when persistent evidence must be transported through frozen features under the measured contract, a small recurrent state supplies the identified gain and an attractive resource profile.

6 Conclusion

MCS replaces a bundled memory win with a controlled question: does information carried across decisions change success when every competing access path is fixed or crossed explicitly? For LSB, the answer is a 4.1-point effect concentrated on S13, TextWorld, and ALFWorld, followed by a 5.4-point advantage over the matched native FLC-4096 configuration and replication with final-layer recurrence. Task-only and trainable-history comparisons then locate that mechanism among alternatives with different parameter, interaction, latency, memory, and training-compute costs. The resulting lesson is actionable: persistent-history control is a regime where bounded temporal

carry can convert frozen language-model features into more accurate decisions without repeatedly processing the retained prefix.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting, editorial revision, and submission-preparation checks. The authors reviewed the AI-assisted work and take responsibility for the technical content, evidence, citations, and final text.

REPRODUCIBILITY STATEMENT

Section 3 specifies the architecture, information-flow contract, objectives, and MCS profile. Section 4 reports training, evaluation, uncertainty, software, and hardware settings; the appendices provide target construction, complete task-wise results, raw denominators, resource measurement details, and integrity checks.

REFERENCES

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11079–11091, 2022. doi: 10.52202/068431-0805. URL https://arxiv.org/abs/2207.06881.

Egor Cherepanov, Nikita Kachaev, Artem Zholus, Alexey K. Kovalev, and Aleksandr I. Panov. Unraveling the complexity of memory in RL agents: An approach for classification and evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=lJKdOYFF5W.

Egor Cherepanov, Aleksei Staroverov, Alexey K. Kovalev, and Aleksandr I. Panov. Recurrent action transformer with memory. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=kByN4v0M3e.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & Miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.13831.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. TextWorld: A learning environment for text-based games. In *Computer Games*, volume 1017 of *Communications in Computer and Information Science*, pp. 41–75. Springer, 2019. doi: 10.1007/978-3-030-24337-1_3. URL https://doi.org/10.1007/978-3-030-24337-1_3.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285/.

Marc Peter Deisenroth and Carl Edward Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 465–472, 2011. URL https://icml.cc/2011/papers/323_icmlpaper.pdf.

Onno Eberhard, Michael Muehlebach, and Claire Vernade. Partially observable reinforcement learning with memory traces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14934–14949. PMLR, 2025. URL https://proceedings.mlr.press/v267/eberhard25a.html.

Kevin Esslinger, Robert Platt, and Christopher Amato. Deep transformer Q-networks for partially observable reinforcement learning. *CoRR*, abs/2206.01078, 2022. doi: 10.48550/arXiv.2206.01078. URL https://arxiv.org/abs/2206.01078.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Matthew Hausknecht and Peter Stone. Deep recurrent Q-learning for partially observable MDPs. In *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, 2015. URL https://arxiv.org/abs/1507.06527.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0RDcd5Axok.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2117–2126. PMLR, 2018. URL https://proceedings.mlr.press/v80/igl18a.html.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Chenhao Lu, Ruizhe Shi, Yuyao Liu, Kaizhe Hu, Simon Shaolei Du, and Huazhe Xu. Rethinking transformers in solving POMDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33089–33112. PMLR, 2024. URL https://proceedings.mlr.press/v235/lu24h.html.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

John Nguyen, Sid Wang, Ke Li, and Carole-Jean Wu. READ: Recurrent adaptation of large transformers. *CoRR*, abs/2305.15348, 2023. doi: 10.48550/arXiv.2305.15348. URL https://arxiv.org/abs/2305.15348.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16691–16723. PMLR, 2022. URL https://proceedings.mlr.press/v162/ni22a.html.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphaël Lopez Kaufman, Aidan Clark, Seb Noury, Matthew Botvinick, Nicolas Heess, and Raia Hadsell. Stabilizing transformers for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7487–7498. PMLR, 2020. URL https://proceedings.mlr.press/v119/parisotto20a.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/arXiv.2307.09288. URL https://arxiv.org/abs/2307.09288.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748. URL https://arxiv.org/abs/1807.03748.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for Bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkl9JlBYvr.

A Protocol and Objective Details

A.1 Auxiliary targets and causal serialization

The recovery head $g_\phi(\mathbf{b}_t)$ concatenates continuous coordinates with softmax probability blocks for categorical factors, while $\mathbf{s}_t$ concatenates normalized continuous targets with the corresponding one-hot vectors. Thus the squared term below is ordinary regression on continuous coordinates and a Brier score on categorical coordinates. The target covers action-relevant simulator factors: for MiniGrid, pose, target identity, carried object, and room-progress bits; for TextWorld, room, quest predicates, and inventory; and for ALFWorld, room, task template, receptacle, object location, task progress, and appliance state. Its recovery loss is

$$\mathcal{L}_{\text{state}} = \frac{1}{T}\sum_{t=1}^{T} \|g_\phi(\mathbf{b}_t) - s_t\|_2^2. \tag{7}$$

The replay buffer stores separate hashes of the normalized current observation, the complete causal observation/action prefix, and the latent target. For each valid anchor, one positive comes from a distinct rollout with the same latent hash, while qualifying negatives in the PPO minibatch match the current-observation hash, differ in latent hash, and differ in complete-prefix hash. Pairs with identical complete-prefix hashes are excluded from the contrastive loss and assigned to the ambiguity-calibration set. Similarity is cosine similarity after ℓ_2 normalization, and every embedding is recomputed by the current forward pass rather than cached across updates. Thus the contrastive term learns state agreement across distinct histories and separates states that are aliased at the current observation without conflicting with calibrated uncertainty for identical complete inputs. Let $\mathcal{V}$ be steps with at least one positive and one observation-matched negative:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \log \frac{\exp(\text{sim}(\mathbf{b}_t, \mathbf{b}_t^+)/\tau_c)}{\exp(\text{sim}(\mathbf{b}_t, \mathbf{b}_t^+)/\tau_c) + \sum_j \exp(\text{sim}(\mathbf{b}_t, \mathbf{b}_{t,j}^-)/\tau_c)}, \qquad \tau_c = 0.1. \quad (8)$$

The term is zero when $\mathcal{V}$ is empty. If $h = (h_1, \ldots, h_K)$ denotes the categorical factor tuple, the categorical blocks induce $\widehat{p}_\phi(h \mid \mathbf{b}_t) \propto \prod_k \widehat{p}_{\phi,k}(h_k \mid \mathbf{b}_t)$, normalized over task-valid hashes; the Brier terms in $\mathcal{L}_{\text{state}}$ therefore train this distribution before entropy regularization. A complete prefix is marked observationally ambiguous when the buffer contains at least two latent hashes for that prefix. On those prefixes, $\mathcal{L}_{\text{calib}} = -T^{-1} \sum_t \mathbb{I}[\text{ambig}_t] H[\widehat{p}_\phi(\cdot \mid \mathbf{b}_t)]$ encourages calibrated uncertainty. All serialization, masks, and recurrent updates are causal at evaluation time.

Table 6: Operational configuration used across rows unless a panel states otherwise.

Component	Exact setting
Transition stream	Literal fields *Task:*, *Observation:*, and *Previous action:*, each followed by its normalized UTF-8 text and a newline; *Previous action: START* at $t = 0$. Whitespace is collapsed before native backbone tokenization, and selected-layer reads use the final non-padding token.
History and reset	LSB encodes the current transition block only. FLC concatenates blocks oldest-to-newest and retains the newest 2048 or 4096 tokens. Recurrent state is zeroed at episode reset and detached between 256-step PPO rollouts.
MiniGrid	MiniGrid 2.3.1 environments *MiniGrid-MemoryS9-v0*, *MemoryS11-v0*, and *MemoryS13-v0*; native caps of 405, 605, and 845 steps; seven native discrete actions. MiniGrid defines no dataset split, so each frozen policy is scored on 200 fresh procedural resets per training seed.
TextWorld	TextWorld 1.6.1 medium quest preset with six rooms, 12 objects, and quest length five. Training uses generator seeds 0–999; held-out game IDs are *tw-1009*, *tw-1013*, *tw-1019*, *tw-1021*, and *tw-1031*, with 40 evaluation episodes per game and seed and a 50-command cap.
ALFWorld	ALFWorld 0.3.5 *eval_out_of_distribution* split; 200 episodes per seed are sampled from its fixed 134-environment evaluation set across six task types, with a 50-action cap.
Action scoring	MiniGrid uses a linear categorical head. TextWorld and ALFWorld form one query from the policy state and score each benchmark-provided admissible command by dot product with its frozen mean native-token embedding; softmax is restricted to that candidate list.
Auxiliary FLC	State recovery and calibration consume the final retained-window token; contrastive similarity uses its normalized projection. Policy and value receive that same token, and no state persists outside the retained window.
GRU readouts	The GPT-2 last-layer comparator uses width 128. The LLaMA cross-workflow rows use width 256; in both cases policy and value receive only recurrent state.
DTQN-style	Six causal transformer blocks, width 512, eight heads, dropout 0.1, and a 2048-token retained window. The encoder and action-query head are trained with task-only PPO on the common online stream; this is an encoder adaptation, not the published Q-learning algorithm.
RATE-style	Eight recurrent transformer blocks, width 512, eight heads, dropout 0.1, 16 memory tokens, and 512-token segments. The encoder and action-query head are trained with task-only PPO; the control adapts segment recurrence but does not implement offline return conditioning or the final RATE Memory Retention Valve.
Selection	All trainable-history controls use the same 128-environment collection interface, admissible-action scoring, optimizer settings, episode caps, and validation-success checkpoint rule as LSB; only the displayed environment-step budgets differ.

A.2 MATCHED-SUPERVISION COMPARISON

Table 7: Broad GPT-2 Medium reference configurations. Success is percentage of solved episodes; uncertainty is one standard deviation over five seeds.

Method	S9	S11	S13	TextW.	ALFW.	Avg
FLC-2048	82.1 ± 2.1	71.5 ± 2.4	58.3 ± 3.2	54.2 ± 2.8	48.7 ± 3.1	63.0
Memory tokens	83.4 ± 2.0	74.8 ± 2.3	61.9 ± 2.9	56.1 ± 2.6	50.8 ± 3.0	65.4
GRU head	85.7 ± 1.8	76.2 ± 2.1	64.1 ± 2.7	58.9 ± 2.5	52.3 ± 2.9	67.4
Selective FT	88.6 ± 1.5	81.0 ± 1.9	72.5 ± 2.4	63.4 ± 2.3	57.1 ± 2.7	72.5
LSB	94.2 ± 1.3	87.3 ± 1.8	79.8 ± 2.2	73.8 ± 2.1	69.5 ± 2.5	80.9

LSB exceeds the P/V-bypass row by 29, 30, and 34 solved episodes per 1,000 on S13, TextWorld, and ALFWorld. The corresponding seed directions are $4/5$, $4/5$, and $5/5$. Matching the auxiliary package therefore preserves the persistent-task ordering while eliminating training-signal availability as its explanation.

Table 8: Full matched-supervision suite on GPT-2 Medium. Every row receives the same frozen backbone, state-recovery, contrastive, and calibration targets, 5M-step PPO budget, and five-seed evaluation protocol. The P/V-bypass row exposes a projected final-layer token to policy and value; the GRU-head row moves recurrent integration to that final frozen layer.

Method	S9	S11	S13	TextWorld	ALFWorld	Avg
LSB	94.2 ± 1.3	87.3 ± 1.8	79.8 ± 2.2	73.8 ± 2.1	69.5 ± 2.5	80.9
LSB + P/V bypass	93.1 ± 1.4	85.6 ± 1.7	76.9 ± 2.4	70.8 ± 2.4	66.1 ± 2.7	78.5
GRU head + auxiliaries	91.0 ± 1.8	83.6 ± 2.0	74.2 ± 2.6	68.0 ± 2.5	63.8 ± 2.8	76.1

B Estimating the Carry Coordinate

Table 9: Complete recurrence-by-bypass factorial under C_G. Every row uses the same augmented head geometry; bypass-off supplies a zero vector of the projected-bypass width.

Carry	Bypass	S9	S11	S13	TextW.	ALFW.	Persist.
On	Off	94.2 ± 1.3	87.3 ± 1.8	79.8 ± 2.2	73.8 ± 2.1	69.5 ± 2.5	74.4
On	On	93.9 ± 1.2	86.4 ± 1.5	77.2 ± 2.0	71.3 ± 2.2	67.0 ± 2.4	71.8
Reset	Off	94.0 ± 1.4	86.9 ± 1.7	75.1 ± 2.6	68.9 ± 2.8	64.0 ± 3.0	69.3
Reset	On	93.6 ± 1.5	86.6 ± 1.8	74.1 ± 2.8	68.0 ± 2.9	63.8 ± 3.1	68.6

Table 10: Solved episodes for the recurrence-by-bypass factorial. Every task contributes 1,000 episodes; "All" pools 5,000 and "Persistent" pools 3,000.

Carry	Bypass	S9	S11	S13	TextW.	ALFW.	All	Persist.
On	Off	942	873	798	738	695	4046	2231
On	On	939	864	772	713	670	3958	2155
Reset	Off	940	869	751	689	640	3889	2080
Reset	On	936	866	741	680	638	3861	2059

The five-task averages for carry/off, carry/on, reset/off, and reset/on are 80.9, 79.2 ± 1.7, 77.8 ± 2.0, and 77.2 ± 2.1. Persistent-average seed standard deviations are 1.9, 2.3, and 2.4 for the three controlled rows. The nonuniformity is the key factorial result. State carry is practically equivalent to reset on S9/S11 under the two-point margin, contributes 4.7–5.5 points on every persistent task without bypass, and remains positive with bypass enabled. The positive interaction follows because restricted decision access increases the persistent carry effect from 3.2 to 5.0 points rather than merely adding an independent offset.

Table 11: Task-wise effect of recurrent carry when bypass is disabled. Deltas are success points; confidence intervals use paired hierarchical resampling.

Task	Carry effect	95% CI	Seed wins
S9	0.2	[-0.8, 1.2]	3/5
S11	0.4	[-0.9, 1.6]	3/5
S13	4.7	[2.7, 6.8]	5/5
TextWorld	4.9	[2.6, 7.1]	4/5
ALFWorld	5.5	[2.9, 8.0]	5/5
Persistent avg.	5.0	[3.6, 6.5]	5/5

C Native-Context and Backbone Controls

Table 12: Legacy GPT-2 retained-context endpoints. FLC-4096 uses the non-native extended prefix; the native-window result in Table 3 carries the inferential comparison.

Setting	FLC-2048	FLC-4096	LSB	Gap vs. 2048	Gap vs. 4096
MiniGrid-S9	82.1	91.4	94.2	12.1	2.8
ALFWorld	48.7	61.2	69.5	20.8	8.3

Table 13: Backbone consistency on the persistent-history subset. The selective-fine-tuning contrast remains positive after moving from GPT-2 Medium to LLaMA-2-7B.

Backbone	Method	S13	TextWorld	ALFWorld	Avg. Δ
GPT-2 Medium	Selective FT	72.5 ± 2.4	63.4 ± 2.3	57.1 ± 2.7	—
	LSB	79.8 ± 2.2	73.8 ± 2.1	69.5 ± 2.5	$+10.0$
LLaMA-2-7B	Selective FT	76.8 ± 2.0	68.9 ± 2.2	62.3 ± 2.5	—
	LSB	83.6 ± 1.7	78.2 ± 1.9	74.1 ± 2.2	$+9.3$

D Label-Free and Retrained-History Comparisons

On S13, task-only LSB records 754 solves, projected bypass 726, and DTQN-style 762 out of 1,000. Across the persistent subset, LSB exceeds bypass by 2.6 points (95% CI [1.1, 4.1]; 4/5 paired seed wins) and differs from DTQN-style by 0.3 points ([-1.3, 1.9]). RATE-style at the matched 5M-step budget reaches 69.3; its 1.4-point advantage over LSB has a 95% interval for LSB minus RATE-style of [-3.0, 0.2].

Table 14: Label-free evaluation on all five tasks. Recurrent rows remove state recovery, contrastive, and calibration losses; external models use observation/action streams only. Means and standard deviations are over five seeds.

Method	S9	S11	S13	TextW.	ALFW.	Persist.
LSB task-only	90.8 ± 1.7	83.7 ± 2.1	75.4 ± 2.5	67.2 ± 2.4	61.0 ± 2.5	67.9
Projected bypass	90.1 ± 1.8	82.8 ± 2.2	72.6 ± 2.8	64.7 ± 2.6	58.6 ± 2.7	65.3
DTQN-style	91.2 ± 1.5	84.0 ± 1.9	76.2 ± 2.3	66.5 ± 2.7	60.1 ± 2.6	67.6
RATE-style, 5M	91.0 ± 1.6	84.2 ± 2.0	77.1 ± 2.4	68.4 ± 2.6	62.3 ± 2.8	69.3

Table 15: Descriptive full-budget operating points. DTQN-style and RATE-style train their history encoders on the same serialized stream, while supervision and budgets remain explicit because neither is matched to the auxiliary LSB row.

Method	Superv.	S9	S11	S13	TextW.	ALFW.	Persist.	Budget
LSB	Aux.	94.2 ± 1.3	87.3 ± 1.8	79.8 ± 2.2	73.8 ± 2.1	69.5 ± 2.5	74.4	5M
DTQN-style	Task	94.4 ± 1.2	88.0 ± 1.6	80.4 ± 1.9	70.9 ± 2.5	64.7 ± 2.6	72.0	10.0M
RATE-style	Task	94.7 ± 1.1	88.5 ± 1.4	82.1 ± 1.7	73.9 ± 2.3	67.7 ± 2.4	74.6	12.0M

E Cross-Workflow Task-Group Replication

Table 16: The complete GRU-head replication. LLaMA residuals use the auxiliary-trained native FLC-4096 comparator; only the task-only GRU rows omit auxiliary supervision.

Backbone	Contract	S9	S11	S13	TextW.	ALFW.	Short	Persist.
GPT-2	GRU + aux.	91.0 ± 1.8	83.6 ± 2.0	74.2 ± 2.6	68.0 ± 2.5	63.8 ± 2.8	87.3	68.7
GPT-2	GRU task-only	85.7 ± 1.8	76.2 ± 2.1	64.1 ± 2.7	58.9 ± 2.5	52.3 ± 2.9	81.0	58.4
LLaMA	GRU + aux.	94.3 ± 1.1	87.8 ± 1.5	81.7 ± 1.9	76.2 ± 2.2	71.7 ± 2.4	91.1	76.5
LLaMA	GRU task-only	93.6 ± 1.3	87.0 ± 1.7	80.4 ± 2.1	74.5 ± 2.4	70.0 ± 2.6	90.3	75.0

The LLaMA GRU + auxiliary row solves 943, 878, 817, 762, and 717 episodes per 1,000. Its residual against auxiliary-trained native FLC-4096 is 0.4 points on the short average (95% CI [-0.6, 1.3]) and 3.3 on the persistent average ([2.0, 4.6]; 4/5 seed wins). Removing auxiliary labels from the GRU reduces its persistent average by 1.6 points ([0.4, 2.8]) yet preserves a positive 1.7-point residual against the auxiliary-trained context carrier.

F Resource Measurements

Resource measurements use BF16 on eight A100 80GB GPUs with 128 environments and 16 environments per GPU. Memory is the maximum allocated device memory after a 500-step warm-up. Decision latency uses batch one, 500 warm-up decisions, and 10,000 timed decisions in each of five repetitions. The FLC measurement re-encodes the retained prefix at each decision and does not reuse a cross-decision KV cache. Throughput times accelerator-resident policy-update steps at the full training batch and excludes environment generation, evaluation, checkpointing, data marshaling, and synchronization. GPU-hours equal eight times end-to-end wall-clock training hours across all five seeds and three persistent tasks, including those excluded costs; uncertainty propagates variation across the five seed-level task totals.

Table 17: Trainable-state and memory measurements. Parameter fraction is relative to the frozen or trainable backbone used by each method; Adam-state memory excludes frozen weights.

Method	Params (M)	Fraction	Adam state (GB)	Train mem. (GB)	Infer mem. (GB)
LLaMA LSB	1.94	0.029%	0.016	39.6 ± 0.8	17.3 ± 0.2
LLaMA FLC-4096	0.58	0.009%	0.005	64.8 ± 1.2	24.7 ± 0.3
LLaMA proj. bypass	2.47	0.037%	0.020	41.2 ± 0.9	17.8 ± 0.2
DTQN-style	47.6	100%	0.381	28.6 ± 0.7	8.4 ± 0.2
RATE-style	72.8	100%	0.582	34.9 ± 0.9	10.9 ± 0.2

Table 18: Sequence, training, and serving measurements. Token counts and throughput report mean $\pm$ standard deviation across five measurement repetitions; GPU-hours report the five-seed, three-task total $\pm$ propagated seed-level standard deviation.

Method	Tokens/decision	GPU-hours	p50 (ms)	p95 (ms)	Steps/s
LLaMA LSB	186 ± 54	612 ± 41	31.8	46.7	438 ± 17
LLaMA FLC-4096	3276 ± 611	1328 ± 86	89.6	132.4	177 ± 9
LLaMA proj. bypass	186 ± 54	657 ± 45	34.1	50.8	411 ± 16
DTQN-style	768 ± 144	1184 ± 72	18.7	27.9	612 ± 24
RATE-style	512 ± 96	1646 ± 103	23.5	35.8	486 ± 20

G Causal Access and Information-Flow Checks

Table 19: Information-flow checks connect each access boundary to a controlled intervention, the required invariant, and its observed evidence.

Boundary	Controlled intervention	Interpretive target	Observed evidence
Causal history	Serialize the current transition and previous carried state only	$\mathbf{b}_t$ contains evidence from the causal prefix only	Equations 1–2 and the rollout reset rule define the complete evaluation path
Policy input	Remove simulator-state targets from recurrent and bypass training	Actions depend only on the observation/action stream	Task-only persistent success is 67.9, with a 2.6-point residual over projected bypass; Table 14
Retained history	Compare bounded recurrence with native FLC-4096	FLC carries no state outside the causal retained prefix	3276 tokens/decision leave a small short residual and a 5.4-point persistent-history residual; Table 3
Latent recovery	Group duplicate prefixes with distinct latent hashes	Recovery remains uncertain for observationally identical prefixes	Entropy calibration regularizes task-valid latent hashes on ambiguous prefixes; Eq. 3
Parity stress	Future offsets 1–8 in training; causal input at evaluation	Sensitivity to train/evaluation access mismatch	Success drops 8.5–14.3 points; Fig. 4

The architecture and serialization define the causal information boundary directly. The deliberately noncausal training control then tests whether future-dependent representations remain useful when evaluated through that boundary. Label-free and native-context comparisons measure carrier utility under the same deployment stream, while entropy calibration handles observational aliasing separately from temporal access.

H Additional Diagnostics

H.1 Bottleneck width

The non-monotonic curve rules out a simple larger-state explanation. Dimensions 64–128 retain enough history for the persistent subset without the optimization degradation observed at larger widths.

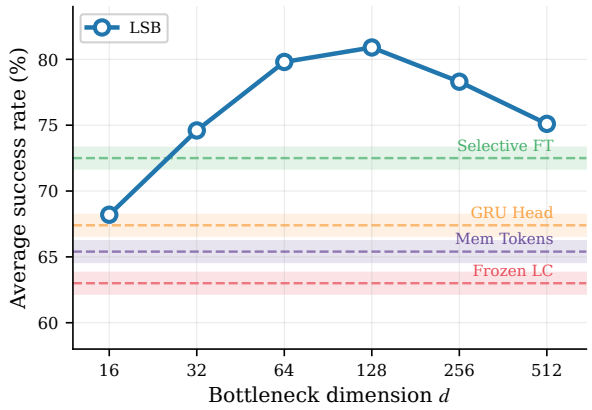


Figure 3: The LSB width sweep has a broad middle optimum. Five-task mean success rises from 68.2 at $d = 16$ to 80.9 at $d = 128$, then declines to 78.3 and 75.1 at $d = 256$ and 512.

H.2 Supervision and access diagnostics

The paired diagnostic separates the utility of contrastive state supervision from the effect of training with an information path unavailable at evaluation.

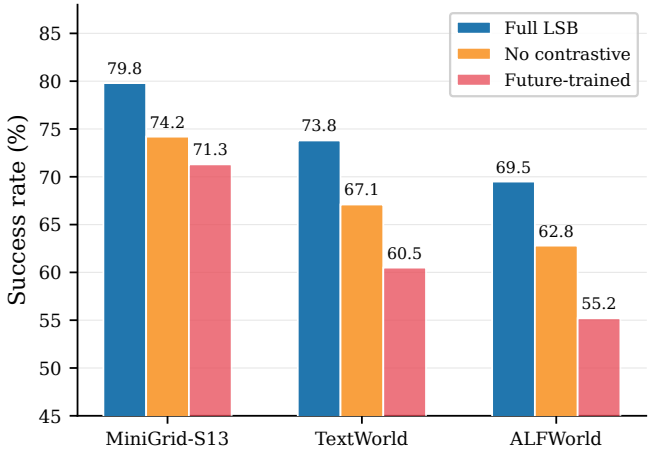


Figure 4: Persistent-history success under the default objective, without contrastive supervision, and after training with future observations at offsets 1–8 but restoring the causal mask for evaluation.

Removing contrastive supervision reduces success by 5.6, 6.7, and 6.7 points on S13, TextWorld, and ALFWorld. Routing observations from offsets 1–8 into the state update during training, then restoring the default causal input at evaluation, produces larger drops of 8.5, 13.3, and 14.3 points. Thus the persistent-history result is strongest when training and evaluation share the causal information path. Standard deviations use the same five seeds as the primary comparisons.