Measuring What Must Persist: A MiniGrid Diagnostic of State-Formation Scaffolding

Anonymous authors
Paper under double-blind review

Abstract

Partially observable RL agents must decide what information should persist as internal state. We treat this as a measurement problem on a controlled MiniGrid (Chevalier-Boisvert et al., 2023) cue_recall diagnostic, using a slot module (ASF) as an instrument: keep/forget gates, cue-decode readouts, and causal memory interventions—not as a general-purpose memory architecture. Unless noted, training-run success is the last 100 episodes of a 50k-step PPO log (3-seed); frozen-policy and distillation numbers are 200-episode evals. With privileged cue-identity losses and event-gated writes, ASF reaches 84.7% vs. 29.7% feedforward PPO and 35.0% GRU; the training auxiliary cue-head decodes the cue from memory at 100%, and zero-masking memory at decision drops success by 21–64pp across privileged Claims A–D. The same template, with delay, distractors, and slot budget varied one at a time, yields 262 selected runs in the submission inventory across four memory pressures. Retention loss (-34.3pp at b=40) and cue-identity auxiliary losses (-11.3pp vs. Claim A full) are load-bearing; the compression loss is not. Removing both cue losses and privileged writes collapses ASF to 3-way chance (30.7%), alongside a retuned GRU (31.3%): the 84.7% result measures scaffolding, not slots versus recurrences. Replacing simulator cue_id with a pixel readout—no probe labels at train or test—yields 90.0% from-scratch ASF. Distilling that student's write vector, with no discrete cue_id at test and no privileged teacher, is within 0.3pp of a 200-episode eval of the student (92.8% vs. 92.5%; n=3, not an equivalence test). Writing observation tokens remains at 3-way chance. The load-bearing object is an identity-shaped write channel; PPO does not invent it (23.0% from-scratch write-head). Under our protocols the identity write channel is required for the observed gains, but its effect is architecture-dependent and high-variance: the same identity write yields $72.7\%_{\pm 13.4}$ (88/63/67) in a Transformer (w=16) vs. 37.0% in a GRU. We report this isolation rather than claiming a new memory module.

1 Introduction

Partially observable sequential decision problems require agents to summarize history into a state that supports later control. Standard RL memory is a fixed architectural choice—a GRU hidden vector, an attention buffer, or a latent world-model state—and success is then read as evidence that the architecture works. We ask a measurement question instead:

What must persist as state—and which training ingredients are load-bearing?

We use a small slot module, Adaptive State Formation (ASF), as an instrument on a MiniGrid (Chevalier-Boisvert et al., 2023) cue_recall unit test: observe a color, wait through b blank steps, then act. The module can be read out and intervened on (auxiliary cue-head on M_t; zero-mask at decision). Privileged cue-identity losses and event-gated writes produce high success; removing both returns ASF to chance, matching a retuned GRU (Figure 1). Pixels can fill the same write channel when they supply cue identity; a write-head trained from sparse reward alone cannot. The result is not that slots beat recurrences

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

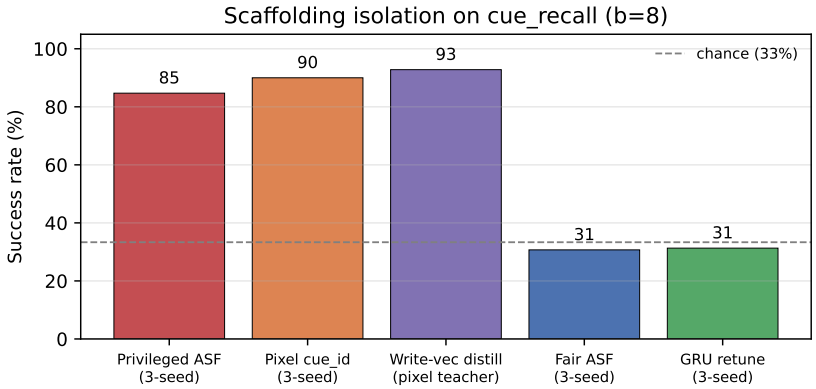


Figure 1: Headline isolation on cue_recall (b=8). Privileged ASF uses cue-identity losses and event-gated writes (Claim A, last-100, 3-seed). Pixel cue_id trains from scratch without simulator cue_id (last-100, 3-seed). Write-vec distill compares that student's EventEncoder from pixels at test (200 episodes, 3-seed). Fair ASF writes observation tokens; GRU is the retuned baseline (both last-100, 3-seed). 3-way chance ∼33%, not 7-action random.

(Section 2). We treat a single diagnostic template as an instrument: the load-bearing object is an identity-shaped write channel whose sufficiency depends on explicit addressable read/write structure (GRU vs. Transformer), not on slot count alone. Our out-of-family smokes (Memory-S, navigation-heavy variants) remain negative (Limitations).

Privileged Claims A–D (delay, distractors, slot budget) characterize the instrument when scaffolding is on; they are not a fair architecture comparison. All main privileged and fair training results use three random seeds. Training-run success is the last 100 logged episodes of 50k-step PPO unless noted; distillation and frozen-policy numbers use 200-episode evals.

Contributions.

- A MiniGrid diagnostic that treats memory as a measurement problem: cue-decode readouts, causal zero-mask, and an explicit privileged vs. fair split.
- Isolation of load-bearing scaffolding: privileged ASF 84.7% vs. fair ASF 30.7% and retuned GRU 31.3% (3-seed, last-100). The GRU stores the cue (held-out blank-phase linear probe 91.4%) but fails delayed control.
- Without simulator cue_id, a pixel student reaches 90.0% and write-vector distillation 92.8%; observation-token and from-scratch write-heads stay at chance (30.7% / 23.0%).
- Identity writes are required under our protocols but architecture-dependent: $72.7\%_{\pm 13.4}$ (Transformer) vs. 37.0% (GRU).
- Mechanism negatives, reported as such: retention loss and predictive-control training are load-bearing; the compression loss and event-gated decoy-control claims are not.

2 Related Work

POMDPs and belief states. Partially observable MDPs formalize decision making under hidden state (Kaelbling et al., 1998). Classical belief-state methods maintain distributions over latent states but scale poorly. We do not learn a belief filter; we measure which write-channel ingredients make a delayed action possible on a unit test.

Recurrent, transformer, and external memory. Recurrent policies encode history in GRU/LSTM hidden states (Hausknecht & Stone, 2015; Mnih et al., 2016). Transformer memory uses bounded attention (Chen et al., 2021; Janner et al., 2021; Parisotto et al., 2020); transformers help memory more than long-horizon credit assignment (Ni et al., 2023).

Addressable writes appear in NTMs, DNCs, and neural episodic control (Graves et al., 2014; 2016; Pritzel et al., 2017). Selective SSMs (Gu & Dao, 2024), Decision Mamba (Huang et al., 2024), world-model SSMs (Samsami et al., 2024), and Decision-RWKV (Dong et al., 2024) revisit linear-time memory; Memory Gym (Pleines et al., 2025) and POPGym Arcade (Wang et al., 2025) extend PopGym-style unit tests. We adopt the circuit-hypothesis premise that memory must be tested by intervention, not probe accuracy alone (Shi et al., 2024); our contribution is the isolation protocol—paired decode readouts and zero-mask interventions that separate architecture limits, missing training signal, and missing write-channel structure. PopGym, Memory Gym, and POPGym Arcade (Morad et al., 2023; Pleines et al., 2025; Wang et al., 2025) report per-architecture success on unit tests but not why agents fail; that decomposition, and its transfer across GRU and Transformer, is the finding—not ASF as a new memory module. We follow this genre on MiniGrid (Chevalier-Boisvert et al., 2023) cue_recall and report official Memory-S as a negative navigation smoke. We do not evaluate Mamba/RWKV backbones; we isolate training-time scaffolding.

Memory-augmented agents and slots. Neural episodic control and MERLIN combine memory with RL for fast value lookup (Pritzel et al., 2017; Wayne et al., 2018). Relational recurrent memory and slot attention offer structured representations (Santoro et al., 2018; Locatello et al., 2020). We use one episodic slot bank on cue_recall (K_s=0); we do not evaluate semantic consolidation or task sequences.

World models. Dreamer-style agents learn latent RSSM states for imagination (Hafner et al., 2021; 2025). We do not evaluate ASF inside a world-model backbone.

3 Problem Formulation

We consider a POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, \gamma)$ where the agent observes $o_t \in \mathcal{O}$, acts $a_t \in \mathcal{A}$, and receives reward r_t. Full history $H_t = (o_{\leq t}, a_{<t}, r_{<t})$ grows unboundedly. Our goal is a persistent state M_t that behaves like a compact sufficient statistic for control and prediction.

State formation map. We define

$$M_t = \mathrm{ASF}_\phi(H_t) = \mathrm{ASF}_\phi(o_{\leq t}, a_{<t}, r_{<t}), \tag{1}$$

with $M_t = \{m_t^1, \ldots, m_t^K\}$ a set of K memory slots. Each slot stores a content vector and gate metadata (importance, age, usage). Unlike a recurrent hidden state, the slot state can be probed and zero-masked; that is why we use it as an instrument, not because the gates are known to be load-bearing.

Desiderata. We say M_t is a good persistent state when:

$$\text{(Compact)} \quad C(M_t) \ll |H_t|, \tag{2}$$

$$\text{(Predictive)} \quad M_t \to \hat{z}_{t+1:t+k}, \hat{r}_{t:t+k}, \hat{e}_{t:t+k}, \tag{3}$$

$$\text{(Control-sufficient)} \quad \pi_\theta(a_t \mid M_t) \text{ achieves high return.} \tag{4}$$

Here z_t denotes task-relevant latent variables (agent pose, inventory, events), and e_t event indicators (key seen, door open).

Training objective. We do not optimize a mutual-information objective. The trained loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{RL}} + \lambda_f \mathcal{L}_{\mathrm{future}} + \lambda_c \mathcal{L}_{\mathrm{compression}} + \lambda_s \mathcal{L}_{\mathrm{stability}}. \tag{5}$$

$\mathcal{L}_{\mathrm{RL}}$ is PPO (Schulman et al., 2017). $\mathcal{L}_{\mathrm{future}}$ supervises prediction of task-relevant futures from M_t rather than raw pixels. $\mathcal{L}_{\mathrm{compression}}$ penalizes active slot usage; $\mathcal{L}_{\mathrm{stability}}$ optionally regularizes slot drift.

Cue decode readout. On cue_recall we report two distinct readouts. For privileged Claims A–D, auxiliary cue-head accuracy applies the same linear head trained with $\mathcal{L}_{\mathrm{cue}}$ to the read context during blank/decision phases—it measures alignment with the training

objective, not a post-hoc probe. Where stated (fair GRU, pixel-cue students), we fit a held-out linear probe on frozen M_t or GRU h_t with train/test episode splits. Neither readout establishes that M_t is a full Markov state.

Compression–control frontier. Privileged Claim D reports success vs. active slots as K varies. We do not interpret that curve as a trained information bottleneck: the compression loss does not bind.

4 Adaptive State Formation (ASF-v1)

ASF is the diagnostic vehicle: a small slot memory whose contents can be probed and masked. Fair comparisons in Section 5.6 show that the gates and slots do not beat a retuned GRU without write-channel scaffolding.

4.1 Observation Tokenization

At each step the agent encodes an observation token

$$x_t = f_\theta(o_t, a_{t-1}, r_{t-1}). \tag{6}$$

Encoders are environment-specific: tile embeddings for MiniGrid. The token dimension is $d = 128$ unless noted otherwise.

4.2 Persistent Memory Structure

ASF maintains two slot banks:

$$M_t = [E_t, S_t], \tag{7}$$

where $E_t \in \mathbb{R}^{K_e \times d}$ are episodic slots and $S_t \in \mathbb{R}^{K_s \times d}$ are semantic slots. We use $K_e = 1$–4 and $K_s = 0$ in every reported run: the semantic bank and abstract update are implemented but unused. Each slot has content, gate value $g_i \in [0, 1]$, age, and usage count. A slot is counted as active when $g_i > \tau$ with τ=0.05 (the code default). Post-hoc thresholding of the same gates at $\tau \in \{0.02, 0.10, 0.20\}$ leaves the Claim D active-ratio trend unchanged (Table 16).

4.3 Keep, Forget, and Abstract

Given token x_t and previous memory M_{t-1}, a gate network outputs soft operation probabilities:

$$(p_{\text{keep}}, p_{\text{abstract}}, p_{\text{forget}}) = \text{softmax}\big(W[x_t, \bar{m}_{t-1}]\big), \tag{8}$$

where $\bar{m}_{t-1}$ is a gated summary of M_{t-1}.

Keep (episodic write). Content-based addressing selects episodic slots; selected slots are updated:

$$m_t^i \leftarrow m_{t-1}^i + p_{\text{keep}}\, \alpha_i\, \Delta_{\text{write}}(x_t, m_{t-1}^i). \tag{9}$$

Forget. Low-value information decays slot gates:

$$g_i \leftarrow g_i \cdot (1 - \eta\, p_{\text{forget}}). \tag{10}$$

Abstract (unused here). The implementation can merge tokens into a semantic bank S_t. With K_s=0 this update is a no-op in all experiments; we do not claim prototype abstraction.

The updated episodic memory is a soft mixture of keep/forget (differentiable gates).

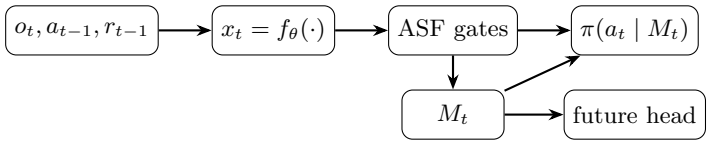


Figure 2: ASF step used as an instrument. Token x_t updates slots M_t through keep/forget (and an unused abstract head). Policy and an optional future head read M_t. Event-gated writes are privileged scaffolding, not a claimed learned gate.

4.4 Reading for Control and Prediction

Policy context is obtained by gated attention over all slots:

$$c_t = \sum_i \mathrm{softmax}(q^\top \tilde{m}_t^i + \log g_i)\, \tilde{m}_t^i, \tag{11}$$

where $\tilde{m}_t^i$ is a learned read projection. The policy and value heads consume c_t; a future-prediction head forecasts task-relevant variables over horizon k.

4.5 Training Losses

Control. MiniGrid uses PPO on the ASF read context (or GRU/Transformer hidden state for baselines).

Future prediction. We predict task-relevant targets, not pixels. MiniGrid targets include agent pose, key/door flags, distance-to-goal, and next reward.

Compression. Version A (default): slot usage penalty $\mathcal{L}_{\text{compression}} = \frac{1}{K}\sum_i g_i$. Optional unused variants include gate entropy and a KL-to-prior term; we do not optimize an information bottleneck.

Cue-specific supervision (MiniGrid cue recall). When the environment exposes cue labels, we add auxiliary heads on the read context c_t:

$$\mathcal{L}_{\text{cue}} = \lambda_w \mathcal{L}_{\text{write}} + \lambda_k \mathcal{L}_{\text{keep}} + \lambda_r \mathcal{L}_{\text{retain}} + \lambda_d \mathcal{L}_{\text{decision}}, \tag{12}$$

where $\mathcal{L}_{\text{write}}$ supervises slot content during cue exposure, $\mathcal{L}_{\text{retain}}$ penalizes drift during blank phases, and $\mathcal{L}_{\text{decision}}$ decodes the cue at the decision step. The shared linear cue head on the read context is therefore a training auxiliary, not an independently fitted post-hoc probe; held-out probes are used only where explicitly stated (fair GRU, pixel-cue students). Default weights: λ_w=1.0, λ_k=0.5, λ_r=0.5, λ_d=0.5.

Event-gated writes. When the environment exposes phase labels (cue visible / blank / distractor), a write policy returns event, retain, or suppress; during retain/suppress, episodic writes are attenuated. This is privileged side information, not a learned gate: we treat it as optional scaffolding and ablate it in Section 5.6. An optional EventEncoder can write probe features (including cue identity) into slots; the fair setting disables this and writes observation tokens only on non-uniform frames, freezing memory on constant blanks. The default token encoder mean-pools cell embeddings over the 7×7 view; a content-cell variant pools only cells that differ from the dominant triple, so walls do not dominate the write token. A pixel cue-id student keeps the EventEncoder but replaces simulator cue_id with a classifier (or center-cell readout), so test-time interaction never reads probe labels. A stricter variant distills the EventEncoder output from pixels on those students' rollouts; test then never constructs probe features and no privileged teacher is involved.

Observation dropout and stability. With probability p=0.15, observation tokens are zeroed during training (regularizer; not strictly necessary at 50k). $\mathcal{L}_{\text{stability}} = \|M_t - M_{t-1}\|_2^2$ prevents slot oscillation but is ablated by default.

4.6 Out of scope

We do not report world-model, Crafter, or multi-task results.

The ASF interaction step (Algorithm 1) is in Appendix A.7.

5 Experiments

We organize experiments as a measurement of which ingredients are load-bearing for persistent state, not as a claim that slots beat recurrences in general. Privileged cue_recall runs (Claims A–D) use cue-identity losses and event-gated writes; Section 5.6 removes both. Within the privileged setting we vary one memory pressure at a time: (A) retention under delay, (B) robustness to extended delays (including OOD generalization), (C) selectivity under distractors, and (D) compression–control trade-offs under slot budgets. These four dimensions are structurally independent—e.g., an agent could retain the cue yet fail to ignore distractors, or succeed at b=8 yet collapse under longer blanks. All runs fit on a single RTX-class GPU; hyperparameters and seeds are in Appendix A.1. Tables and figures are auto-generated from the 262-run inventory plus 26 follow-up runs (Appendix 4); fair and distillation rows are curated separately (Appendix A.6).

5.1 Baselines and Metrics

Task: cue_recall. The agent observes a colored cue for T_{cue}=3 steps, then a blank corridor of length b steps (optionally with distractors or decoy cues), and must turn toward the matching door at the end. Partial view size is 7×7. The MiniGrid action space has 7 discrete actions. A uniform-random policy scores 13.8%$_{\pm 3.1}$ (200 episodes × 3 seeds: 16/9.5/16), near 1/7. PPO and GRU instead converge to 30–35%, close to the 3-way conditional chance among $\{\text{left}, \text{right}, \text{forward}\}$ (∼33%) rather than that 7-action floor (Table 5): they learn to restrict to the three task-relevant actions but fail to condition the choice on the cue. We call ∼33% 3-way chance.

Baselines. PPO (feedforward on o_t), PPO+GRU (128-d hidden state), and PPO+Transformer (bounded attention window matched to ASF slot count in Claim D). All agents share the PPO optimizer (lr 2.5×10^{-4}, rollout 256, 50k steps). We did not run a full learning-rate grid for PPO+GRU or PPO+Transformer (Table 2); numbers are a same-budget comparison, not a best-effort ceiling. A minimal check at the Claim A GRU default (h=128, full BPTT, b=8) with lr $\in \{1\times10^{-4}, 5\times10^{-4}\}$ (3-seed each) stays at 31.3%$_{\pm 4.2}$ and 34.3%$_{\pm 2.6}$, within 4pp of the default 35.0%$_{\pm 7.8}$; we did not sweep hidden width. Privileged Claim A–D GRUs use full BPTT over the 256-step PPO trajectory; fair GRUs use TBPTT 16 (Table 15). Claim A–D ASF numbers use privileged writes unless noted.

Metrics. (1) episode success rate; (2) active memory ratio $|\{i : g_i > \tau\}|/K$ with τ=0.05 (Table 1; training default). Recomputing the ratio at $\tau \in \{0.02, 0.10, 0.20\}$ on saved Claim A/D checkpoints does not change the K=1→4 drop (Table 16); (3) cue decode accuracy—auxiliary cue-head on the read context for privileged ASF (Claims A–D), or held-out linear probe on frozen states where stated; (4) causal intervention: zero-mask M_t at decision and measure success drop; (5) compression–control Pareto across slot budgets K.

Reporting protocol. Figure 1 is the headline isolation; Claims A–D are privileged (not a fair slots-vs-GRU comparison). Last-100 success is the mean of the final 100 training episodes in a 50k-step log; distillation and frozen-policy numbers are 200-episode greedy evals. Captions state which protocol is used.

5.2 Claim A: Write and Retain (privileged)

Table 5 (Appendix A.6) and Figure 3 report the privileged anchor (b=8, 50k, 3 seeds). Privileged ASF reaches 84.7% vs. 29.7% (PPO) and 35.0% (GRU); GRU h=256 is 33.2%—capacity alone does not close the gap (Table 15). Across all three seeds, the auxiliary

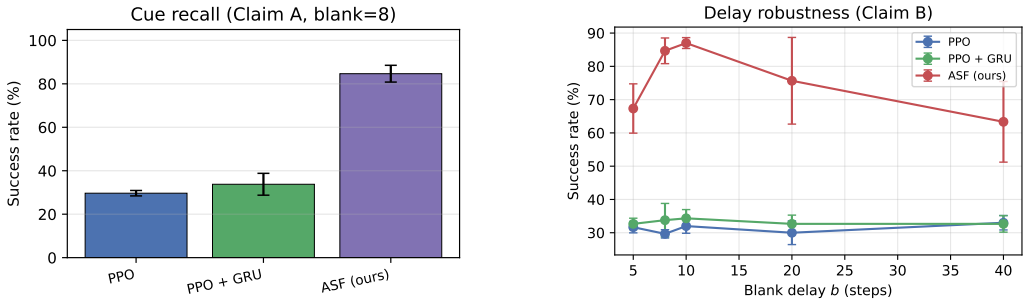


Figure 3: Left: Privileged Claim A success by baseline (3 seeds). Right: Privileged Claim B delay sweep ($b \in \{5, 8, 10, 20, 40\}$). PPO/GRU remain near chance; event-gated ASF degrades more slowly.

cue-head decodes the cue from M_t at 100% (Tables 5 and 12); zero-masking at decision drops success by 51–64pp on Claim A (3-seed min–max 51.0/64.0/63.5; Table 5 reports 60 ± 6 as mean $\pm$ population SD). Section 5.6 shows this gap does not survive without privileged writes.

5.3 Claim B: Delay Robustness (privileged)

Table 6 sweeps blank delay $b \in \{5, 8, 10, 20, 40\}$ with training steps scaled as $50000 \times b/8$ to match episode budget. PPO and GRU stay flat at ∼30–35% (near 3-way chance). Those Claim B GRU numbers use full-episode BPTT, not the fair TBPTT-16 recipe (Table 15): at b=40 GRU is $32.7\%_{\pm 2.5}$ (32/30/36), so the delay gap is not a 16-step TBPTT artifact (Table 8). Privileged ASF maintains 63.3% at b=40 (+30.3pp over PPO) and 84.7% at the b=8 anchor (Table 6). Active slot usage remains stable (∼0.92) across delays (full table in Appendix A).

Extreme-delay stress test. Separately, we train at b=60 and b=100 with K_e=4 (Table 7); these are not a continuation of the K_e=1 curve. PPO and GRU stay near 3-way chance at b=60 (29.7% and $30.0\%_{\pm 3.7}$). At b=100, PPO last-100 success is $27.0\%_{\pm 4.1}$ (32/22/27; seed 2 recovered after rerun). GRU is $31/29/0^{\dagger}$ per seed (Table 7): seed 2 collapsed to a constant-action policy (supplementary investigation log), which we treat as a training failure—ASF seed 2 still reaches 33%, ruling out an environment bug. Two supplementary GRU runs (seeds 3–4, same 625k schedule) reach 32% and 35%, so the collapse is seed-specific rather than a universal long-delay baseline failure; we therefore report per-seed values instead of a 3-seed aggregate for GRU at b=100. Privileged ASF reaches 88.7% (b=60) and 44.3% (b=100).

5.4 Claim C: Distractor Selectivity (privileged)

During the blank phase, $n \in \{0, 3, 5, 8, 12\}$ colored distractor tiles appear; only the original cue determines reward. Figure 4 and Table 9 (Appendix A) show privileged ASF success flat at 82–88% while baselines stay at 30–38%. At n=12, privileged ASF achieves 87.7% vs. GRU 33.7% (+54pp).

5.5 Claim D: Compression–Control Pareto (privileged)

We sweep episodic slot count $K \in \{1, 2, 4\}$ for privileged ASF (Table 10, Appendix A). ASF at K=1 reaches 84.7% with active ratio 0.92; at K=4 success rises to 87.0% while active slots drop to 0.46. GRU and Transformer remain near 3-way chance at the K we trained (Table 10: GRU is the Claim A default 35.0 ± 7.8; Transformer w=K is 34.0 ± 1.4 / 33.3 ± 2.5 / 30.0 ± 4.2 at K=1, 2, 4).

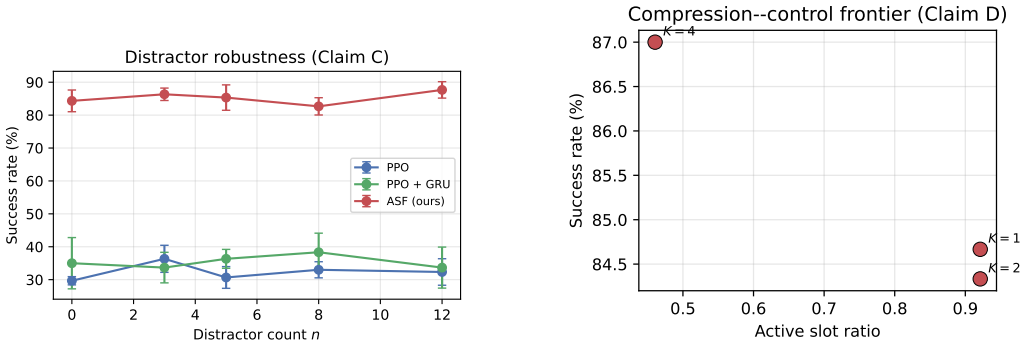


Figure 4: Left: Privileged Claim C distractor density sweep (3 seeds). Right: Privileged Claim D compression–control frontier at $K \in \{1, 2, 4\}$.

5.6 Fair comparison: measuring scaffolding

The main cue_recall results use a privileged write-channel pipeline: cue-identity auxiliary losses, event-gated writes, an EventEncoder over simulator features (including cue_id), content-cell pooling, and blank-phase write freezing. The fair setting removes this entire pipeline—not two isolated factors—and writes observation tokens only (Figure 1, Table 11 in Appendix A.6).

Fair protocol. Observation-token writes only (no EventEncoder); all λ_{cue}=0; writes occur iff the image is not spatially uniform (blank MiniGrid frames repeat a single cell—no cue_seen / cue_id). A retuned GRU (MLP policy head, truncated BPTT of 16, which still covers the b=8 episode of length 12; Table 15) and a GRU with cue-classification aux are trained under the same 50k-step PPO budget. Unlike Claim B's b-scaled training budget ($50000 \times b/8$), the fair GRU extension at $b \in \{20, 40\}$ keeps that same 50k-step budget, isolating TBPTT-16 shrinking relative to episode length from extra training steps (Table 8). Success remains at 3-way chance: 32.3%$_{\pm 3.1}$ at b=20 and 29.7%$_{\pm 1.2}$ at b=40.

Result. Privileged ASF (Claim A) is 84.7% (3-seed); predictive-only ASF (cue losses off, privileged writes kept) stays at 73.3%. Fair content-write ASF is 30.7%$_{\pm 2.1}$ (3-seed: 33/28/31); always-write (30.7%$_{\pm 2.5}$), content-cell (30.3%$_{\pm 2.5}$), a retuned GRU (31.3%$_{\pm 1.2}$), GRU+aux (35.0%$_{\pm 1.7}$), and retain/decision losses without EventEncoder writes (35.0%$_{\pm 4.6}$) all stay at 3-way chance (n=3; not an equivalence test). The two 35.0% GRU point estimates are different configs (Table 15). The architecture without scaffolding does not leave chance; the privileged 84.7% is therefore a measurement of labels and write gates, not of slots versus recurrences.

GRU is not an empty encoder. A linear probe on frozen hidden states of the retuned 50k GRU reaches 91.4% cue accuracy on held-out blank-phase episodes: the GRU stores the cue across b=8 blanks, so the 3-way-chance success is a control-mapping failure. A Markov control with b=0 yields 67% GRU success at 50k—the same PPO stack leaves chance when delay is a single decision frame. Injecting the same identity write channel reaches 37.0%$_{\pm 5.0}$ in the retuned GRU and 72.7%$_{\pm 13.4}$ in a w=16 Transformer (high seed variance). Gains require both that channel and explicit read structure that exposes it to the policy; sufficiency is architecture-dependent.

Pixel and write-vector distillation. Fair agents fail because they never map pixels into the privileged write channel; content-cell pooling does not rescue fair writes (30.3%$_{\pm 2.5}$, Table 11). A pixel cue-id student without simulator cue_id reaches 90.0%$_{\pm 6.5}$; write-vector distillation at test is 92.8%$_{\pm 3.7}$ vs. 92.5% (0.3pp, n=3, not an equivalence test). A from-scratch write-head stays at 23.0%$_{\pm 8.7}$ (Table 11).

5.7 Mechanism Ablations (Stage 2)

Stage 2 ablates ASF mechanisms on b=8 (Table 12): 3-seed Claim A anchors plus seed 0 dagger rows not compared to the 84.7% mean. Predictive only drops to 73.3% ($-$**11.3**pp; probe 21%); at b=40, no_cue_retain drops $-$34.3pp. Other toggles change success by $\leq$8pp; decoy event-gating does not replicate (Section B; Table 13). Sample efficiency tracks the same trend (Table 14). Privileged ASF exceeds baselines on A–D; Section 5.6 shows this does not survive scaffolding removal.

6 Limitations

Diagnostic scope. Main evidence is the cue_recall family (delay, distractors, budget, OOD delay). That isolates write-channel ingredients; it does not substitute for multi-environment benchmarking. OOD delay (b=60, 100) uses K_e=4 and is not a continuation of the K_e=1 curve.

Mechanism ablations. Core mechanism tests and all fair variants are 3-seed; ancillary auxiliary toggles remain seed 0. The compression loss is a confirmed negative (K=8, budget 0.04: 89.3% with λ_c=0.05 vs. 89.7% with λ_c=0; probe 100% both; Table 3). Decoy event-gating is withdrawn (Section B). All runs use soft gates with K_s=0. Pixel distillation shapes identity before fair PPO (Section 5.6); a from-scratch write-head stays at 23.0%$_{\pm 8.7}$. Identity-channel injection is architecture-dependent (72.7%$_{\pm 13.4}$ Transformer vs. 37.0%$_{\pm 5.0}$ GRU; Section 5.6).

Environment generalization. Outside the cue_recall family we report smokes only, not full claims. Official MiniGrid Memory-S is a T-maze without cue_id, so the ASF write pipeline never fires; a 15k seed-0 run gives GRU 40% vs. ASF 6% (retuned TBPTT-16; Table 15). A hallway/split-intervention diagnostic on the same seed-0 checkpoints is under development but not yet stable. Unlike cue_recall, Memory-S has variable episode length and no single synchronized decision step, so clearing hidden state or slots whenever $x \geq 5$ fires triggers over multiple timesteps with episode-dependent timing. Paired with $\sim$40% success, this makes the induced success drop swing from -2 to $+6$ pp across reruns (binomial noise on 200 episodes); we do not report it until the protocol is fixed. spatial_memory, delayed_cue (privileged ASF 45%), and cue_recall_multi (5% vs. fair 3%) are likewise single-seed navigation probes, not a second domain.

7 Conclusion

We treated persistent memory in RL as a measurement problem; ASF is the instrument—slots with cue-decode readouts and zero-mask interventions—not a claimed general-purpose module. Without the privileged write-channel pipeline, ASF sits with a retuned GRU at 3-way chance (30.7% / 31.3%, 262 selected runs, 3-seed); with it, privileged ASF is 84.7%. Pixels supply the missing write-channel identity (90.0% student; 92.8% write-vector distill, no simulator cue_id ever), which PPO cannot invent (23.0%). The 84.7% privileged number measures write-channel identity plus explicit read structure, not slots versus recurrences. The practical takeaway is that control-sufficient memory in delayed POMDPs may depend more on how information enters addressable state than on whether that state is slotted, recurrent, or attention-based—a hypothesis our protocol is meant to export, not a MiniGrid-specific trick.

AI use statement

Generative AI tools were used to aid and polish writing (including LaTeX formatting and code/documentation scaffolding) and for literature organization (retrieval and discovery). They were not used to draft paper sections or to invent experimental results. Bibliography metadata (authors, titles, venues) for all cited works was checked against official proceedings or arXiv pages after an earlier pass had mixed author lists across DeepMind papers; remaining errors are the authors' responsibility. Claim A–D tables and figures are auto-generated

via scripts/export_paper_results.py from the 262-run submission inventory (Appendix 4) plus the 26-run follow-up batch in runs/followup/; fair-comparison, pixel/write-vector, and distillation rows are curated separately (Table 11). We take responsibility for the final content.

Reproducibility statement

Code, configs, and training scripts are provided in the supplementary repository. Section 4 and Algorithm 1 specify ASF-v1; Appendix A.1 lists hyperparameters and seeds. MiniGrid environments (Appendix A.2) export simulator true state for Markov probes and event-gated write labels. Experiment configs are YAML files under configs/minigrid/. Regenerate paper tables and figures with:

python scripts/export_paper_results.py

We will release checkpoints and full metric logs with the camera-ready version. Analyses we did not run—full hidden-width baseline HPO and formal equivalence tests—are correspondingly absent from the released scripts (Limitations); a minimal GRU lr spot-check at b=8 is in Section 5.1. Headline comparisons (84.7% vs. 30.7%/31.3%) exceed cross-seed variance (∼4pp SD) by an order of magnitude and do not hinge on formal testing. We do not report Welch / TOST tests for the numerically close comparisons in Section 5.6 (e.g., 92.8% vs. 92.5%): n=3 seeds give those tests too little power to be informative, so we report per-seed values instead.

References

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In Advances in Neural Information Processing Systems (NeurIPS), 2021.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, 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 (NeurIPS), 2023.

Yujian Dong, Tianyu Wu, and Chaoyang Song. Optimizing robotic manipulation with decision-RWKV: A recurrent sequence modeling approach for lifelong learning, 2024.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. arXiv preprint arXiv:1410.5401, 2014.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 538:471–476, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In First Conference on Language Modeling (COLM), 2024.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In International Conference on Learning Representations (ICLR), 2021.

Danijar Hafner, Jurgis Pašukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. Nature, 2025.

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.

Sili Huang, Jifeng Hu, Zhejian Yang, Liwei Yang, Tao Luo, Hechang Chen, Lichao Sun, and Bo Yang. Decision mamba: Reinforcement learning via hybrid selective sequence modeling. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In Advances in Neural Information Processing Systems (NeurIPS), 2021.

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.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Proceedings of the 33rd International Conference on Machine Learning, pp. 1928–1937, 2016.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. PopGym: Benchmarking partially observable reinforcement learning. In International Conference on Learning Representations (ICLR), 2023.

Tianwei Ni, Michel Ma, Benjamin Eysenbach, and Pierre-Luc Bacon. When do transformers shine in RL? decoupling memory from credit assignment. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

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, pp. 7487–7498, 2020.

Marco Pleines, Matthias Pallasch, Frank Zimmer, and Mike Preuss. Memory gym: Towards endless tasks to benchmark memory capabilities of agents. Journal of Machine Learning Research, 26(6):1–40, 2025.

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

Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. Mastering memory tasks with world models. In International Conference on Learning Representations (ICLR), 2024.

Adam Santoro, Ryan Faulkner, David Raposo, Jack W. Rae, Mike Chrzanowski, Théophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational recurrent neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in LLMs. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

Zekang Wang, Zhe He, Borong Zhang, Edan Toledo, and Steven Morad. Investigating memory in RL with POPGym arcade, 2025.

Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwińska, Jack Rae, Piotr Mirowski, Joel Z. Leibo, Adam Santoro, Mevlana Gemici, Malcolm Reynolds, Tim Harley, Josh Abramson, Shakir Mohamed, Danilo Rezende, David Saxton, Adam Cain, Chloe Hillier, David Silver, Koray Kavukcuoglu, Matt Botvinick, Demis Hassabis, and Timothy Lillicrap. Unsupervised predictive memory in a goal-directed agent. arXiv preprint arXiv:1803.10760, 2018.

A Supplementary Material

A.1 Implementation Details

Setting	Value
Token / slot dimension d	128
Episodic slots K_e (main)	1 (budget sweep: 1/2/4)
Semantic slots K_s	0 (cue recall)
Gate hidden dim	256
Future horizon k	5
λ_f (future)	0.05
λ_c (compression)	0.01
$\lambda_w, \lambda_k, \lambda_r, \lambda_d$	1.0, 0.5, 0.5, 0.5
PPO rollout steps	256
Learning rate	2.5×10^{-4}
Total training steps	50k ($\times b/8$ for delay sweep)
View size	7×7
Seeds	3 (0, 1, 2)
Eval episodes (diagnostics)	200
Active-gate threshold τ	0.05 (training default; post-hoc Table 16)
Privileged GRU/Transformer BPTT	full 256-step PPO unroll (tbptt_steps=0)
Fair GRU TBPTT	16 (covers b=8 episode length 12)
Reported $\pm$	population SD across seeds (ddof=0)

Table 1: Default ASF-v1 and training hyperparameters for cue_recall. All $\pm$ in this paper are population SD across seeds unless a caption quotes min–max.

Baseline	Success (%)
PPO+GRU (h∈{128, 256}, TBPTT {0, 16}, MLP, cue-aux; lr spot-check at b=8)	35.0±7.8 (default lr)
PPO+Transformer (w∈{1, 2, 4}, Claim D w=K)	34.0±1.4 / 33.3±2.5 / 30.0±4.2

Table 2: Baseline tuning actually performed. We did not run a full lr grid; Section 5.1 reports a 3-seed spot-check at lr $\in \{10^{-4}, 5\times10^{-4}\}$. Transformer numbers are w=1/2/4 from Table 10.

A.2 Environment Details

cue_recall. Cue visible for T_{cue}=3 steps; blank corridor of length b; agent must turn toward the door matching cue color. 3-way cue identity maps onto {left, right, forward}; MiniGrid still exposes 7 actions. Uniform random over 7 scores 13.8%$_{\pm 3.1}$ (200 episodes $\times$ 3 seeds); 3-way chance among {left, right, forward} is ∼33%. PPO/GRU sit at the latter.

cue_recall_distractor. Same as above, but n colored distractor tiles appear each blank step; reward uses the original cue only.

cue_recall_decoy. During blank, decoy cue colors flash every 2 steps; reward always refers to the first cue seen. Tests whether event-gated writes prevent overwrite.

Each environment exports probe labels: cue color, phase (cue/blank/decision), distractor flag, and memory_retain indicator.

A.3 Diagnostic Protocol

After training, we evaluate 200 episodes with:

- Baseline: normal rollouts; auxiliary cue-head accuracy on the read context (the linear head trained with $\mathcal{L}_{\text{cue}}$—not a post-hoc probe). Held-out linear probes on frozen states are reported separately for fair GRU and pixel-cue students.
- Zero-mask: set all slot gates to 0 at decision step.
- Shuffle: permute slot contents at decision (sanity check).
- Wrong-cue: inject incorrect cue into memory (sanity check).

Success drop under zero-mask is reported in percentage points (pp).

A.4 Ablation Details

Ablation	Modification	Observed effect
Predictive only (n=3)	$\mathcal{L}_{\text{cue}}$=0	−11.3pp unrounded (−11.4pp from 84.7−73.3); cue-head 21%
No retention (n=3, b=40)	λ_r=0	−34.3pp vs. full (3-seed replicated)
No retention, decoy (n=3)	λ_r=0	aux. cue-head −30.3pp
No write supervision (seed 0)	λ_w=0	no asymptotic gap
With λ_c=0.05 (K=8, tight)	loss on	89.3 (95/90/83); probe 100%
Without λ_c=0 (K=8, tight)	loss off	89.7 (92/89/88); probe 100%; Δgate 0.002
No event gating, decoy (n=3)	disable probe gating	probe ≈0pp (seed-0 only, withdrawn)
No obs. dropout (seed 0)	p=0	+4pp success
No forget (seed 0)	η=0	≤3pp on distractor/decoy

Table 3: Stage 2 ablation protocol and measured outcomes. Core mechanisms (retention, predictive-control, compression) are 3-seed; ancillary auxiliary ablations remain seed 0.

A.5 Run Inventory

Directory	Runs	Seeds	Content
runs/stage1_v2/	18	3	Claim A
runs/stage1_claim_b/	63	3	Claim B delay sweep ($b\in[5,40]$)
runs/stage1_claim_c_v2/	45	3	Claim C distractor
runs/stage1_claim_d/	27	3	Claim D budget
runs/stage2_ablation/	52	0–3	Mechanism ablations
runs/stage3_blank/	12	3	Claim B OOD (b=60, 100)
runs/poster_fair/	45	3	Fair isolation, pixel/write-vec, cross-arch
Total	262		

Table 4: Submission inventory: 262 runs with results.json under the seven directories above. The full runs/ tree has 338 files. The other 76 are: legacy Stage 1 (26), stage1_legacy (12), Claim B v2 draft (9), Claim C v1 (6), Crafter/DQN smokes (18), Memory-S (2), and short Stage 3 delay smokes (3). They are superseded protocols, out-of-scope environments, or smokes whose negatives (Memory-S, Crafter) appear in Limitations rather than the claim tables. Inclusion is by directory membership, not by inspecting success: the 262 submission runs live only under the seven directories above (stage3_blank/ holds PPO and ASF OOD only, 12 runs). Memory-S is reported but not counted in the 262. Post-submission follow-ups live in runs/followup/ (26 runs: fair GRU $b\in\{20,40\}$ [6], lr spot-checks [6], Stage 3 OOD GRU $b\in\{60,100\}$ [8; b=100 includes supplementary seeds 3–4], PPO b=100 reruns [3], Transformer w=2 [3]; verified by find runs/followup -name results.json excluding archive subdirs) and are merged into generated claim tables but excluded from the 262 inventory.

A.6 Supporting Tables

Tables 5–14 in this appendix are the full versions of the results summarized in the main text. Table 15 is the unique index of every GRU configuration named in the paper (including Memory-S); Table 8 records the Claim B GRU BPTT protocol.

Method	Success (%)	Active slots	Probe (%)	Zero-mask Δ (pp)
Uniform random (7 actions)	13.8±3.1	0.00	–	–
PPO	29.7±1.2	0.00	–	–
PPO + GRU	35.0±7.8	0.00	0.0	1±6
ASF (ours)	84.7±3.9	0.92	100.0	60±6
PPO + GRU (h=256)	33.2±2.5	0.00	–	–

Table 5: Claim A: cue_recall with blank=8 (3 seeds). Random row: uniform over 7 MiniGrid actions, 200 episodes × 3 seeds (16/9.5/16). Probe column: auxiliary cue-head on read context (ASF is 100% on all three seeds, hence no ±). Zero-mask Δ is mean ± population SD (pp). ASF seed drops are 51.0/64.0/63.5 (min–max 51–64).

Method	$b = 5$	$b = 8$	$b = 10$	$b = 20$	$b = 40$
PPO	31.7±1.7	29.7±1.2	32.0±2.2	30.0±3.6	33.0±2.2
PPO + GRU	32.7±1.7	33.8±5.0	34.3±2.6	32.7±2.6	32.7±2.5
ASF (ours)	67.3±7.4	84.7±3.9	87.0±1.6	75.7±13.0	63.3±12.1

Table 6: Claim B: success vs. blank delay with K_e=1 (3 seeds). Training steps scale as $50000 \times b/8$. Extreme delays $b \in \{60, 100\}$ use a different slot budget and are reported separately in Table 7.

Method	$b = 60$	$b = 100$
PPO	29.7±3.3	27.0±4.1
PPO + GRU	30.0±3.7	31/29/0†
ASF (ours)	88.7±4.7	44.3±10.3

Table 7: Extreme-delay stress test ($b \in \{60, 100\}$). Unlike Table 6, ASF uses K_e=4. Every agent at these delays uses the Stage 3 schedule: 375k steps at b=60 and 625k at b=100 ($50000 \times b/8$). These points are not a continuation of the K_e=1 delay curve. †GRU b=100: per-seed last-100 success on the main inventory (seeds 0/1/2); seed 2 collapsed to a constant-action policy in both training runs (action 5 only; ASF seed 2 remains 33%); per-seed values reported instead of a 3-seed aggregate. Supplementary GRU seeds 3/4: 32%/35%. PPO seed 2 recovered after rerun (32/22/27).

A.7 Algorithm

A.8 Out of scope

Not in this submission: Crafter, world models, PopGym suites, and lifelong task sequences. We do not claim architecture-agnostic results.

A.9 Reproducibility

All experiments use fixed seeds (0, 1, 2), deterministic PyTorch operations with manual seed control, and auto-generated tables and figures from runs/ via scripts/export_paper_results.py. The full codebase (dependency list, config files, and the 262-run submission inventory in JSON format) will be released upon acceptance. Training any single run requires one command (python scripts/train_minigrid.py) on a commodity desktop GPU; the complete claim matrix can be reproduced within days on a single RTX-class GPU.

Setting	BPTT	b=20	b=40	b=60	b=100
Claim B GRU (Table 6)	full 256-step unroll	—	32.7 ± 2.5 (32/30/36)	—	—
Fair retuned GRU (50k)	TBPTT 16	32.3 ± 3.1 (34/35/28)	29.7 ± 1.2 (30/28/31)	—	—
Stage 3 OOD GRU (Table 7)	full BPTT	—	—	30.0 ± 3.7 (29/26/35)	31/29/0[†]

Table 8: Credit-assignment window vs. delay. Claim B GRU uses full-episode BPTT; at b=40 it remains at 3-way chance. Fair GRU at $b \in \{20, 40\}$ keeps the 50k-step b=8 budget (Table 8). Stage 3 GRU uses the 375k/625k budgets of Table 7; supplementary b=100 seeds 3/4: 32%/35%.

Method	$n = 0$	$n = 3$	$n = 5$	$n = 8$	$n = 12$
PPO	$29.7_{\pm 1.2}$	$36.3_{\pm 4.1}$	$30.7_{\pm 3.3}$	$33.0_{\pm 2.4}$	$32.3_{\pm 4.0}$
PPO + GRU	$35.0_{\pm 7.8}$	$33.7_{\pm 4.6}$	$36.3_{\pm 2.9}$	$38.3_{\pm 5.8}$	$33.7_{\pm 6.2}$
ASF (ours)	$84.3_{\pm 3.3}$	$86.3_{\pm 1.9}$	$85.3_{\pm 3.9}$	$82.7_{\pm 2.6}$	$87.7_{\pm 2.5}$

Table 9: Claim C: distractor density sweep during blank phase (3 seeds).

K	ASF success (%)	Active slots	GRU (%)	Transformer w=K (%)
1	$84.7_{\pm 3.9}$	0.92	$35.0_{\pm 7.8}$	$34.0_{\pm 1.4}$
2	$84.3_{\pm 3.4}$	0.92	$35.0_{\pm 7.8}$	$33.3_{\pm 2.5}$
4	$87.0_{\pm 0.8}$	0.46	$35.0_{\pm 7.8}$	$30.0_{\pm 4.2}$

Table 10: Claim D: memory-budget sweep (K episodic slots for ASF). GRU is the Claim A default (h=128), not re-tuned per K. Transformer window is w=K; w=16 is a separate near-chance smoke (35.7 ± 1.9).

Setting	Success (%)	n	Priv. writes / cue loss
Privileged ASF (Claim A)	$84.7_{\pm 3.9}$	3	yes / yes
Pixel cue_id → frozen teacher	$86.2_{\pm 4.4}$	3	pixel / distilled
Pixel cue_id student (from scratch)	$90.0_{\pm 6.5}$	3	pixel / yes
Write-vector distill (pixel teacher)	$92.8_{\pm 3.7}$	3	pixel vec / no
Write-vector distill (privileged teacher)	$74.5_{\pm 6.0}$	3	pixel vec / no
Write-vector student (from scratch, +vec aux)	$23.0_{\pm 8.7}$	3	pixel vec / no
Predictive-only ASF	$73.3_{\pm 1.2}$	3	yes / no
Fair ASF (content write)	$30.7_{\pm 2.1}$	3	no / no
Fair ASF + content-cell pool	$30.3_{\pm 2.5}$	3	no / no
Fair ASF + retain/decision loss	$35.0_{\pm 4.6}$	3	no / yes
Fair ASF (always write)	$30.7_{\pm 2.5}$	3	no / no
Fair retuned (b=8)	$31.3_{\pm 1.2}$	3	— / no
+ identity channel (b=8)	$37.0_{\pm 5.0}$	3	event / yes
Fair + cue aux (b=8)	$35.0_{\pm 1.7}$	3	— / yes
Transformer + identity write channel	$72.7_{\pm 13.4}$	3	event / yes

Table 11: Scaffolding isolation on cue_recall b=8. Last-100 of 50k PPO except distillation/frozen rows (200-episode eval). Pixel cue_id replaces simulator cue_id (from-scratch, 3-seed). Write-vector distill matches on-policy EventEncoder outputs from pixels. GRU rows use the same labels as Table 15. Unconditional 7-action random is $13.8\%_{\pm 3.1}$; 3-way chance ∼33% (Section 5.1). ± is population SD across seeds.

Algorithm 1 ASF-v1 interaction step

Require: Token encoder f_θ, memory M_{t-1}, gates $\{g_i\}$
1: $x_t \leftarrow f_\theta(o_t, a_{t-1}, r_{t-1})$
2: $(p_k, p_a, p_f) \leftarrow \text{GateNet}(x_t, \bar{m}_{t-1})$
3: Update episodic slots E_t via Keep with weight p_k
4: Abstract update is a no-op (K_s=0)
5: Decay gates via Forget with weight p_f
6: $c_t \leftarrow \text{Read}(M_t)$; sample $a_t \sim \pi(\cdot \mid c_t)$
7: Update θ, ϕ with $\mathcal{L}_{\text{RL}} + \lambda_f \mathcal{L}_{\text{future}} + \lambda_c \mathcal{L}_{\text{compression}}$
8: return a_t, M_t

Variant	Success (%)	Probe (%)	Zero-mask Δ (pp)
Full ASF	84.7±3.9	100.0	59.5±6.0
No cue aux. (pred. only)	73.3±1.2	21.0±15.2	52.0±1.9
No write supervision†	94.0	100.0	60.0
No decision decode†	90.0	100.0	66.5
No obs. dropout†	98.0	100.0	61.0
No compression (K=1)†	90.0	100.0	59.5

Table 12: Stage 2 mechanism ablations on cue_recall blank=8. Top block: 3-seed Claim A anchors (Table 5); Full ASF probe is 100% on all three seeds. †Seed 0 only on the stage2 re-anchor (Full ASF 94% at 50k on that run; Claim A seed 0 is 83%). The dagger band sits with that re-anchor, not above Claim A's 84.7% mean—do not compare magnitudes to the 3-seed Full row.

Variant	Success (%)	Probe (%)	Zero-mask Δ (pp)
Full ASF	21	85.3	-2
No event gating	21	87.7	10
No forget decay	20	100.0	7
No retention loss	21	55.0	5

Table 13: Decoy overwrite stress test: decoy colors appear during blank; reward uses first cue (3 seeds; mean over available seeds).

Variant	10k	25k	50k
Full ASF	35	55	94
No write supervision	36	56	94
No cue aux. (pred. only)	42	45	75

Table 14: Sample efficiency on cue_recall blank=8. Seed 0 stage2 re-anchor (Full 94% at 50k); not Claim A's 3-seed 84.7%.

GRU variant	Ret.	Aux.	ID	BPTT / success (%)
Claim A default (b=8)	no	no	no	full; 35.0 ± 7.8 (34/26/45)
Claim A h=256 (b=8)	no	no	no	full; 33.2 ± 2.5
Claim B delay ($b \leq 40$)	no	no	no	full; 32.7 ± 2.5 at b=40
Fair retuned (b=8)	MLP	no	no	TBPTT 16; 31.3 ± 1.2 (33/31/30)
Fair + cue aux (b=8)	MLP	yes	no	TBPTT 16; 35.0 ± 1.7 (33/36/36)
+ identity channel (b=8)	MLP	no	yes	TBPTT 16; 37.0 ± 5.0 (32/37/42)
Stage 3 OOD (b=60/100)	no	no	no	full; 30.0 ± 3.7 at b=60 (29/26/35); 31/29/0† at b=100
Memory-S smoke (15k, s0)	MLP	no	no	TBPTT 16; 40%

Table 15: Every GRU configuration named in this paper. cue_recall rows are 3-seed last-100 unless per-seed values are given. Stage 3 OOD GRU: follow-up runs under runs/followup/; supplementary b=100 seeds 3/4 are 32%/35% (Table 7). Memory-S is a 15k seed-0 smoke. The two 35.0% point estimates differ in BPTT, variance, and auxiliary loss.

Setting	τ=0.02	τ=0.05	τ=0.10	τ=0.20
Claim A K=1	0.92	0.92	0.92	0.92
Claim D K=2	0.92	0.92	0.92	0.92
Claim D K=4	0.46	0.46	0.46	0.42

Table 16: Active-slot ratio recomputed from Claim A/D checkpoints (200 eval episodes × 3 seeds; no retraining). Gates saturate near 0/1, so the K=1→4 drop is stable across $\tau \in \{0.02, 0.05, 0.10, 0.20\}$. The only movement is K=4 at τ=0.20 (0.46→0.42), and it is consistent across seeds (0.413/0.417/0.417), not a single-seed artifact.

B Analysis

The contribution is the isolation, not a new memory architecture. The diagnostic tests a short hypothesis list (arrows give outcomes):

- slots without scaffolding would beat GRU if sufficient → both at chance;
- cue losses alone would drive control if privileged writes were unnecessary → 73.3% remains without them;
- retain/decision losses on observation tokens would leave chance if supervision alone were enough → 35.0%;
- pixels would recover control if the EventEncoder channel were sufficient → 90.0% from-scratch student;
- discrete cue_id would be required at test if write-vector matching failed → 92.8% (pixel-trained teachers) / 74.5% (privileged);
- PPO would invent the write geometry if a teacher were unnecessary → 23.0% from-scratch write-head;
- the channel alone would be sufficient for any architecture → GRU 37.0% vs. Transformer $72.7\%_{\pm 13.4}$ (high variance);
- the compression loss would bind if gate usage depended on λ_c → it does not.

Cue decode readouts. Privileged Claims A–C report auxiliary cue-head accuracy on the read context (the head used in $\mathcal{L}_{\text{cue}}$), reaching 100% on every 3-seed diagnostic; removing cue auxiliary losses (predictive only) collapses that readout to 21% (3-seed) despite 73.3% success. Fair GRU uses a separate held-out linear probe on frozen h_t (91.4% during blank); the 3-way-chance success is therefore not an empty encoder. Pixel-cue students are the complementary cell: a held-out probe on frozen slots is 100% and zero-masking drops success by 62–75pp while the training cue-head stays near chance.

Causal intervention. At evaluation we zero-mask all memory slots at the decision step. For privileged ASF across Claims A–D (3-seed inventory) this drops success by 21–64 percentage points, confirming that control depends on M_t rather than the instantaneous frame.

Slot budget, not compression loss. Figure 4 (right) plots success vs. active slot ratio across $K \in \{1, 2, 4\}$. Increasing K from 1 to 4 raises success marginally (84.7%→87.0%) while halving active slot usage (0.92→0.46). That frontier is a slot-count effect (Claim D). The compression loss does not bind (Section 6). The insensitivity to τ across $[0.02, 0.20]$ is consistent with gates saturating near 0 or 1 rather than clustering near the threshold (Table 16): active ratio is therefore a near-binary count of engaged slots, not a soft occupancy measure, which is why a default of τ=0.05 is not a hidden degree of freedom in this range.

Event-gated writes. At 3 seeds the decoy event-gating probe advantage does not replicate (87.7% vs. 85.3%, $\Delta = -0.4$pp); the seed-0-only signal is withdrawn. Removing the retention loss causes a consistent probe drop (-30.3pp), indicating that retention—not event gating—protects decodable memory content under overwrite pressure.