GAMPO: AUDITING AND OPTIMIZING LSTM MEMORY FOR PARTIALLY OBSERVABLE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recurrent policies must preserve useful evidence while limiting sensitivity to nuisance history. Forget-gate products omit paths through coupled LSTM cell–hidden dynamics. We use the complete joint-state Jacobian to bound finite history perturbations and define a full-path memory tail. Filter contraction and decoder calibration translate this tail into belief and control errors; a terminal-decision refinement avoids unnecessary horizon factors in recall tasks. Gate-Aware Memory Policy Optimization (GAMPO) combines an exponential teacher with sensitivity, information-state, and projection losses. Its population regret decomposition separates representation, critic, and projection errors, with explicit conditions linking fitted losses to these terms. Four controlled POMDPs reveal substantial limitations: the reported model-free variant remains near chance on recall tasks and can underperform unregularized sequence baselines. Full-path influence improves over a forget-only proxy on active sensing, and an independent fast-refresh control shows a regret benefit over omission. Short-delay and exact-belief controls expose persistent learning limitations even when memory or representation demands are reduced. These findings support full-path sensitivity as an auditing tool and expose the additional requirements for effective memory regularization.

1 INTRODUCTION

An early observation can determine the correct action many decisions later. In a partially observable Markov decision process (POMDP), a recurrent policy must retain useful evidence while suppressing nuisance variation. A small history derivative can indicate robustness or the loss of the only useful cue. Memory analysis therefore requires both a measure of retained influence and a check that the representation supports control.

LSTM memory is not determined by forget gates alone (Hochreiter & Schmidhuber, 1997; Can et al., 2020; Livi, 2026). Hidden-state feedback can change later gates and re-enter the cell state. Our starting point is the complete cell–hidden Jacobian. The perturbation argument applies to differentiable recurrences generally; the gate-level certificates here are specific to LSTMs. We do not claim the joint-state Jacobian or contraction-based finite-context approximation as new: these build on Can et al. (2020); Miller & Hardt (2019). Temporal Range summarizes local derivatives by a magnitude-weighted lag (Lafuente-Mercado et al., 2025). Our focus is a uniform tail controlling finite perturbations and the conditions that turn it into a control bound.

Beliefs are sufficient for POMDP control (Kaelbling et al., 1998). Filter-stability and approximate-information-state analyses connect representation quality to planning loss (McDonald & Yüksel, 2020; Kara & Yüksel, 2022; Subramanian et al., 2022; Golowich et al., 2023). We combine these ideas with recurrent input-entry and state-transport bounds. The implication is conditional: small sensitivity does not establish accurate beliefs. Decoder calibration and environmental filter stability must be justified separately. For tasks with only a terminal decision, a sharper value comparison uses the common predecision history distribution.

Gate-Aware Memory Policy Optimization (GAMPO) uses an exponential teacher inspired by MPO (Abdolmaleki et al., 2018) and an actor-to-teacher KL projection motivated by mirror descent. This projection has the opposite KL orientation to the original MPO M-step. Sensitivity, information-state,

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.

and log-ratio penalties are auxiliary objectives. GAMPO-KF uses a known Bayes filter; the reported GAMPO-P uses finite-lag predictive supervision (Venkatraman et al., 2017; Hefny et al., 2018; Wang et al., 2023). The nominal block loss excludes the terminal reward, but the executed code extends its reward targets through the terminal stage. We also define GAMPO-P-R with a separate terminal-state prediction head; this is distinct from the implemented block extension. Existing RL results concern GAMPO-P and GAMPO-KF; no performance result is claimed for GAMPO-P-R.

Recurrent RL provides strong empirical baselines (Hausknecht & Stone, 2015; Ni et al., 2022; Morad et al., 2023). Our controls use LSTM, GRU, causal Transformer, and S4D encoders (Cho et al., 2014; Vaswani et al., 2017; Gu et al., 2022) within one optimization pipeline. They isolate encoder effects but do not replace independently validated RL algorithms. The analysis complements finite-time recurrent policy-gradient guarantees (Cayci & Eryilmaz, 2025) and policy mirror descent with function approximation (Sherman et al., 2025); it does not prove that nonconvex training achieves its error conditions. Table 15 compares the scope of the closest work.

The contributions are (i) a full-path perturbation-to-control analysis with explicit calibration and filtering requirements; (ii) a gate-aware projection framework and terminal-reward correction to its predictive variant; and (iii) a conditional online decomposition with critic-risk, coverage, and projection-transfer bridges. Controlled experiments and analytic audits identify useful diagnostics and failure modes. Their strongest conclusion concerns the inadequacy of forget-only summaries; overall algorithmic superiority is not established.

2 Problem and Recurrent Information State

We consider an episodic POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{O}, \mathcal{A}, H, \mu, P, \Omega, \mathsf{R})$, with finite action set $\mathcal{A}$, horizon H, standard Borel state and observation spaces, and measurable transition, observation, and reward kernels. In episode i, $S_{i,1} \sim \mu$; at stage h, observation $Y_{i,h} \sim \Omega_h(\cdot \mid S_{i,h})$ precedes action $A_{i,h}$, reward $R_{i,h} \sim \mathsf{R}_h(\cdot \mid S_{i,h}, A_{i,h}) \in [0,1]$, and transition $S_{i,h+1} \sim P_h(\cdot \mid S_{i,h}, A_{i,h})$. The pre-action history and cumulative regret are

$$H_{i,h} = (Y_{i,1}, A_{i,1}, R_{i,1}, \ldots, A_{i,h-1}, R_{i,h-1}, Y_{i,h}),$$

$$\mathrm{Regret}_N(\pi^\star) = \sum_{i=1}^{N} (V^{\pi^\star} - V^{\pi_i}),$$

where $V^\pi = \mathbb{E}_{\pi,\mathcal{M}}[\sum_{h=1}^{H} R_h]$, R_h is a realized reward with episode index suppressed, and $\mathcal{H}_h$ is the admissible history space. Policies may change between episodes using past data. Stage-dependent masks specify nonempty admissible sets $\mathcal{A}_h(H_h) \subseteq \mathcal{A}$. All action sums, maxima, KL divergences, and probability floors below are restricted to these sets; inadmissible actions have probability zero. A forced-action stage has zero policy divergence and advantage oscillation.

The exact belief is $b_{i,h} = \mathbb{P}(S_{i,h} \in \cdot \mid H_{i,h})$. For $h \geq 2$, the Bayes operator incorporating the previous action and reward and current observation gives $b_{i,h} = \Phi_{h,A_{i,h-1},R_{i,h-1},Y_{i,h}}(b_{i,h-1})$. GAMPO-KF uses finite $\mathcal{S}$, probability-vector beliefs, and known Bayes updates. General-state filtering needs a suitable belief representation; the history-space regret analysis permits general $\mathcal{S}$.

To summarize the observable history, encode each tuple as

$$X_{i,h} = \phi(Y_{i,h}, A_{i,h-1}, R_{i,h-1}) \in \mathbb{R}^{d_x},$$

with fixed start symbols at $h = 1$. A width-d_u LSTM (Hochreiter & Schmidhuber, 1997) computes

$$\begin{aligned} I_h &= \sigma(W_I X_h + U_I U_{h-1} + b_I), & F_h &= \sigma(W_F X_h + U_F U_{h-1} + b_F), \\ O_h &= \sigma(W_O X_h + U_O U_{h-1} + b_O), & G_h &= \tanh(W_G X_h + U_G U_{h-1} + b_G), \\ C_h &= F_h \odot C_{h-1} + I_h \odot G_h, & U_h &= O_h \odot \tanh(C_h). \end{aligned} \tag{1}$$

We fix $Z_0 = (C_0, U_0)$ and write the complete update as

$$Z_h = (C_h, U_h) = \mathcal{F}_\theta(Z_{h-1}, X_h).$$

The main results fix ϕ and Z_0; Section C.3 gives the additional terms for learned components.

A decoder $D_\psi : \mathbb{R}^{d_u} \to \mathcal{B} \subseteq \mathcal{P}(\mathcal{S})$ maps the exposed hidden state, rather than the cell state, to an approximate belief,

$$\hat{b}_h = D_\psi(U_h), \qquad \pi_{\vartheta,h}(a \mid H_h) = \bar{\pi}_{w,h}(a \mid \hat{b}_h),$$

where $\vartheta = (\theta, \psi, w)$. The output gate thus controls downstream access to cell memory. GAMPO-KF uses this belief decoder; GAMPO-P and GAMPO-P-R use a predictive state $z_h = D_\psi(U_h) \in \mathbb{R}^{d_D}$.

We call z_h *predictively sufficient* if histories with the same z_h induce the same joint law of future observations and rewards under every admissible fixed future action sequence. This observable condition need not identify latent beliefs. Applying the belief-level results to either predictive variant additionally requires a calibrated map $B_h(z_h) \in \mathcal{B}$, a policy head factoring through this map, and the stated Lipschitz and filtering conditions. Section C.3 formalizes these requirements; finite-lag predictive fitting alone does not establish them.

Global bounds use declared compact recurrent and input domains $\mathcal{Z}, \mathcal{X}$ and parameter set Θ. Jacobian suprema cover all comparison segments. The filtering domain $\mathcal{B}$ contains reachable and decoded beliefs and is invariant under the realized Bayes updates. Equal encoded suffixes must determine equal raw tuples for encoded-history filtering claims. This losslessness condition holds for injective one-hot/affine encoders; learned encoders need a separate sufficiency check (Section C.3). Sampled Jacobians do not certify these global conditions.

Figure 1 summarizes the POMDP interaction and recurrent actor. The LSTM state comprises both cell and hidden coordinates; only the hidden coordinates are exposed to the decoder.

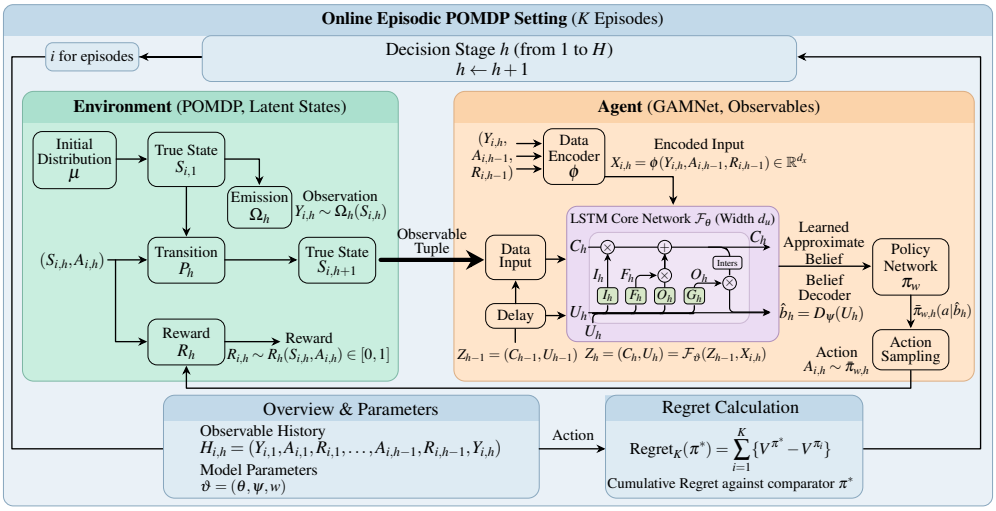


Figure 1: Recurrent POMDP interaction. GAMPO-KF decodes beliefs; GAMPO-P decodes predictive states. The recurrent transition labeled $\mathcal{F}_\vartheta$ in the diagram denotes $\mathcal{F}_\theta$: $Z_h = \mathcal{F}_\theta(Z_{h-1}, X_{i,h})$ depends only on the recurrent parameters θ, while ψ and w parameterize the decoder and policy head. Here K counts episodes (N_K in our batched notation), and the reward kernel R_h is denoted R_h in the text.

3 Full-Jacobian Memory and Sensitivity

3.1 History influence and effective memory

Suppress episode indices and let $D(v) = \mathrm{diag}(v)$. Differentiating Equation (1) gives

$$\begin{aligned} J_h^C &= D(C_{h-1})D\big(F_h \odot (1 - F_h)\big)U_F + D(G_h)D\big(I_h \odot (1 - I_h)\big)U_I \\ &\quad + D(I_h)D(1 - G_h^2)U_G, \\ J_h^U &= D\big(\tanh(C_h)\big)D\big(O_h \odot (1 - O_h)\big)U_O + D(O_h)D\big(1 - \tanh^2(C_h)\big)J_h^C. \end{aligned}$$

For $Z_h = (C_h, U_h)$, the complete recurrent Jacobian is

$$J_h^{ZZ} = \frac{\partial Z_h}{\partial Z_{h-1}} = \begin{bmatrix} D(F_h) & J_h^C \\ D(O_h)D\big(1 - \tanh^2(C_h)\big)D(F_h) & J_h^U \end{bmatrix}. \tag{2}$$

The upper-left block is the direct retention path. The remaining blocks couple cell and hidden perturbations and are absent from a forget-only product. Let $J_h^{ZX} = \partial \mathcal{F}_\theta(Z_{h-1}, X_h)/\partial X_h$. Over the comparison domain described in Section 2, define

$$\alpha_h = \sup \|J_h^{ZZ}\|_{\mathrm{op}}, \qquad \beta_h = \sup \|J_h^{ZX}\|_{\mathrm{op}}, \qquad \Gamma_{h,s} = \beta_s \prod_{j=s+1}^{h} \alpha_j.$$

Proposition 1 (Full-path perturbation)**.** *For two admissible input histories processed by the same recurrent parameters,*

$$\|Z_h - Z'_h\|_2 \le \left(\prod_{j=1}^{h} \alpha_j \right) \|Z_0 - Z'_0\|_2 + \sum_{s=1}^{h} \Gamma_{h,s} \|X_s - X'_s\|_2.$$

The domain-wide suprema cover the interpolation segments, so the bound applies to finite perturbations.

Definition 2 (Memory tail)**.** *For a recent window of m inputs and $\epsilon > 0$, define*

$$T_h(m) = \sum_{s=1}^{\max\{h-m,0\}} \Gamma_{h,s}, \qquad m_{\mathrm{eff}}(h;\epsilon) = \min\{m \in \{0,\dots,h\} : D_X T_h(m) \le \epsilon\},$$

where $D_X = \sup_{x,x' \in \mathcal{X}} \|x - x'\|_2$. If two histories have the same m most recent inputs and the same initial state, then $\|Z_h - Z'_h\|_2 \le D_X T_h(m)$.

A cue separated from decision h by D recurrent transitions is among the retained inputs only when the input window has length at least $D + 1$.

Corollary 3 (Uniform contraction)**.** *If $0 < \bar{\alpha} < 1$, $\alpha_h \le \bar{\alpha}$, $\beta_h \le \bar{\beta}$, and $D_X \bar{\beta} > 0$, then*

$$\Gamma_{h,s} \le \bar{\beta} \bar{\alpha}^{h-s}, \qquad T_h(m) \le \frac{\bar{\beta} \bar{\alpha}^m}{1 - \bar{\alpha}}.$$

Contraction is sufficient, not necessary; norm products may be loose under rotation or cancellation. The explicit window bound is in Section C.1.

Readout derivatives are bounded above by the state-transport envelope, but a small output derivative need not imply a small recurrent-state perturbation. Section C.1 explains this one-way relation and why the implemented penalty remains a local surrogate.

3.2 Parameter sensitivity and statistical scope

For a critic coordinate $f_{\vartheta,h}(H_h, a)$, let

$$K_{\mathrm{net}}^{Q} = \sup_{\vartheta \in \mathrm{co}(\Theta),\, h, H_h, a} \|\nabla_\vartheta f_{\vartheta,h}(H_h, a)\|_2 \le K_{\mathrm{net}}^{Q,+}.$$

This class-uniform certificate includes the recurrent core, decoder, and readout. It controls parameter perturbations; the memory tail instead controls input perturbations for a fixed model. Section C.2 and Proposition 12 give explicit gate-level envelopes. With parameter diameter R and dimension p, these yield sequential complexity of order $RK_{\mathrm{net}}^{Q,+}\sqrt{p \log(n)/n}$ (Rakhlin et al., 2015). Proposition 16 additionally requires independent episodes, bounded returns, approximation, and optimization control to bound critic risk. Sampled gradients do not certify these class-uniform conditions.

4 From Memory to Belief and Control Error

To connect sensitivity to retained information, we combine the memory tail with filtering and calibration conditions.

Assumption 4 (Filter contraction domain)**.** *There is a set $\mathcal{B}$ containing every exact reachable and decoded belief. For every conditioning tuple $\tau = (h, a, r, y)$ used below and every $b \in \mathcal{B}$, the Bayes normalizer is positive and $\Phi_\tau(b) \in \mathcal{B}$. Moreover, for some $\rho_B \in [0, 1)$ and all $b, b' \in \mathcal{B}$,*

$$\mathrm{TV}\{\Phi_{h,a,r,y}(b), \Phi_{h,a,r,y}(b')\} \le \rho_B \, \mathrm{TV}(b, b').$$

This environmental property is distinct from contraction of the learned recurrence. Define $e_1 = \mathrm{TV}(\hat{b}_1, b_1)$ and, for $h = 2, \ldots, H$, the one-step defect

$$\delta_h = \mathrm{TV}\left[D_\psi(U_h), \Phi_{h,A_{h-1},R_{h-1},Y_h}\left(D_\psi(U_{h-1})\right)\right]. \tag{3}$$

Theorem 5 (Filtering-error propagation)**.** *Under Assumption 4,*

$$\mathrm{TV}(\hat{b}_h, b_h) \le \rho_B^{h-1} e_1 + \sum_{s=2}^{h} \rho_B^{h-s} \delta_s.$$

In particular, if $\delta_s \le \bar{\delta}$, the second term is at most $\bar{\delta}/(1-\rho_B)$.

Strict contraction can fail with persistent latent states and uninformative observations. Under nonexpansiveness alone, $\mathrm{TV}(\hat{b}_h, b_h) \le \min\{1, e_1 + \sum_{s=2}^{h} \delta_s\}$, without geometric forgetting. E3 satisfies strict contraction; delayed recall and active sensing do not. Memory perturbation and history-space regret bounds are independent of this assumption (Section C.4).

For each history, assume an admissible representative with the same last m raw tuples and a fixed earlier prefix. Suppose its exact belief differs by at most $C_B \rho_B^m$, its decoder calibration error is at most ϵ_{dec}, and $\mathrm{TV}\{D_\psi(u), D_\psi(u')\} \le L_{D,\mathrm{TV}} \|u - u'\|_2$. Here TV is $\sup_E |b(E) - b'(E)|$ over measurable sets. For n_S finite states, TV is half the ℓ_1 distance, so one may choose $L_{D,\mathrm{TV}} = \sqrt{n_S} L_{D,2}/2$ on the comparison segments; the two decoder constants are distinct. Proposition 1 then yields:

Corollary 6 (Memory-to-belief approximation)**.** *For every h and m,*

$$\mathrm{TV}(\hat{b}_h, b_h) \le \epsilon_{\mathrm{dec}} + C_B \rho_B^m + L_{D,\mathrm{TV}} D_X T_h(m).$$

Calibration is an independent requirement on the representative histories. Reducing $T_h(m)$ may increase calibration error when useful evidence is erased; the bound does not justify minimizing sensitivity alone.

Assumption 7 (Lipschitz policy head)**.** *For every stage and all $b, b' \in \mathcal{B}$,*

$$\mathrm{TV}\{\bar{\pi}_{w,h}(\cdot \mid b), \bar{\pi}_{w,h}(\cdot \mid b')\} \le L_\pi \, \mathrm{TV}(b, b').$$

Let π^b apply a fixed head to the exact belief and let $\hat{\pi}$ apply the same head to $\hat{b}_h$. Write $d_h^\pi = \mathcal{L}_\pi(H_h)$ and $\bar{d}_h^\pi = \mathcal{L}_\pi(S_h, H_h)$, where $\mathcal{L}_\pi$ denotes the probability law induced by policy π and the fixed POMDP. Thus d_h^π is the history marginal and $\bar{d}_h^\pi$ is the joint state–history marginal at stage h.

Theorem 8 (Control-error propagation)**.** *Under Assumption 7, if $e_h = \sup_{H_h} \mathrm{TV}(\hat{b}_h, b_h)$, then*

$$\sup_{H_h} \mathrm{TV}\{\hat{\pi}_h(\cdot \mid H_h), \pi_h^b(\cdot \mid H_h)\} \le L_\pi e_h, \tag{4}$$

$$\mathrm{TV}(\bar{d}_h^{\hat{\pi}}, \bar{d}_h^{\pi^b}) \le L_\pi \sum_{s=1}^{h-1} e_s, \tag{5}$$

$$|V^{\hat{\pi}} - V^{\pi^b}| \le L_\pi \sum_{s=1}^{H} (H - s + 1) e_s. \tag{6}$$

Corollary 9 (Terminal-decision refinement)**.** *Suppose preterminal actions are forced, $R_h = 0$ for $h < H$, and $R_H \in [0,1]$. Under Assumption 7, both policies share a terminal history law d_H, and*

$$|V^{\hat{\pi}} - V^{\pi^b}| \le \mathbb{E}_{d_H} \min\{1, L_\pi \, \mathrm{TV}(\hat{b}_H, b_H)\} \le \min\{1, L_\pi \mathbb{E}_{d_H} \, \mathrm{TV}(\hat{b}_H, b_H)\}. \tag{7}$$

This applies to E1–E3 when the decoder and head conditions hold, without filter contraction. Theorem 8 and Corollary 9 enter comparator approximation separately from the mirror-descent telescope.

5 Gate-Aware Memory Policy Optimization

At policy epoch k, the deployed policy π_k supplies m_k fresh episodes, split into disjoint critic and actor sets $\mathcal{D}_k^Q, \mathcal{D}_k^A$. Actor and critic use separate recurrent cores. For class-uniform results, fix

parameter sets $\Theta^r(G_r)$, $r \in \{A, Q\}$, inside prespecified convex envelopes on which $K^{r,+}_{\text{net}} \le G_r$ is verified. The feasibility wrapper in Section D.3 distinguishes this class-level condition from checks at individual iterates. The reported runs remain diagnostic-only because their training code does not implement this certification wrapper.

In the reference formulation, a full-support actor supplies Monte Carlo returns on $\mathcal{D}^Q_k$ for a clipped recurrent critic. The fit includes a local parameter-sensitivity penalty and is restricted to $\Theta^Q(G_Q)$; its exact normalized objective is in Section D.2. Centering the fitted $\widehat{Q}_k$ under π_k gives $\widehat{A}_{k,h}$. Clipping $\widehat{Q}_{k,h} \in [0, H-h+1]$ ensures $\operatorname{osc}_a(\widehat{A}_{k,h}) \le H-h+1$. Accounting for all m_k deployments, the exponential teacher is

$$\widetilde{\pi}_{k+1,h}(a \mid H_h) = \frac{\pi_{k,h}(a \mid H_h) e^{\eta m_k \widehat{A}_{k,h}(H_h,a)}}{\sum_{a'} \pi_{k,h}(a' \mid H_h) e^{\eta m_k \widehat{A}_{k,h}(H_h,a')}}. \tag{8}$$

GAMPO-KF uses initial-belief and known-filter consistency losses. GAMPO-P predicts finite-lag observation–reward targets. The nominal block indexing excludes R_H; the executed code extends reward blocks through H. A distinct terminal-state variant, GAMPO-P-R, uses

$$\widehat{\mathcal{L}}^{\text{P-R}}_{B,k} = \widehat{\mathcal{L}}^{\text{P}}_{B,k} + \lambda_{\text{T}} \widehat{\mathcal{L}}_{\text{T},k}, \qquad \widehat{\mathcal{L}}_{\text{T},k} = -\frac{1}{|\mathcal{D}^A_k|} \sum_i \log p_{\chi,\text{T}}(R_{i,H} \mid z_{i,H}, A_{i,H}),$$

with prespecified $\lambda_{\text{T}} > 0$. This supplies a direct terminal reward target and backpropagates through the full actor recurrence. Section D.2 gives the losses and a recall-specific identification result. Existing tables do not evaluate GAMPO-P-R.

For the fitted actor's logits ℓ^A_h, we compute local full-path influence $\widehat{J}^A_{i,h,s} = \|\partial \ell^A_h / \partial X_{i,s}\|_{\text{op}}$ and parameter sensitivity $\widehat{K}^{A,\text{loc}}_{i,h} = \|\partial \ell^A_h / \partial \vartheta^A\|_{\text{op}}$. Their normalized squared averages define $\widehat{\mathcal{R}}^A_{\Gamma,k}$ and $\widehat{\mathcal{R}}^A_{K,k}$. A sampled worst-history log-ratio penalty $\widehat{\mathcal{R}}_{D,k} = \widehat{\xi}^2_k$ measures distortion between the teacher and the fitted actor before exploration mixing; exact formulas, lag weights, and numerical approximations are in Section D.2. The square emphasizes large sampled distortions. The regret bound remains linear in the deployed-policy population distortion ξ_k; a sampled loss needs a separate uniform gap bound, plus the explicit exploration correction below, to control that term.

With n_{pow} power iterations, independent traversal of all source–target paths costs $O(n_{\text{pow}} H^3 c_{\text{step}})$ per trajectory; sampling L lags per target gives $O(n_{\text{pow}} L H^2 c_{\text{step}})$. Here c_{step} is one derivative-product step. Matrix-free serial recomputation permits storage linear in H, but adds runtime (Section D.3).

In predictive variants, prediction-head parameters χ are jointly optimized over a declared class; the display shows the actor block. The ideal actor objective is

$$\begin{aligned}\vartheta^A_{k+1} \in \arg\min_{\vartheta \in \Theta^A(G_A)} & \frac{1}{|\mathcal{D}^A_k| H} \Big[\sum_{i,h} \text{KL}\{\pi_{\vartheta,h}(\cdot \mid H_{i,h}) \| \widetilde{\pi}_{k+1,h}(\cdot \mid H_{i,h})\} \Big] \\ & + \lambda_{\Gamma,k} \widehat{\mathcal{R}}^A_{\Gamma,k} + \lambda_{K,k} \widehat{\mathcal{R}}^A_{K,k} + \lambda_{B,k} \widehat{\mathcal{L}}_{B,k} + \lambda_{D,k} \widehat{\mathcal{R}}_{D,k}. \end{aligned} \tag{9}$$

The optimizer defines the fitted base actor $\bar{\pi}_{k+1} := \pi_{\vartheta^A_{k+1}}$. The policy actually deployed in the next epoch is

$$\pi_{k+1,h} = (1 - \gamma_{k+1}) \bar{\pi}_{k+1,h} + \gamma_{k+1} u_h, \tag{10}$$

where u_h is uniform over admissible actions. This notation separates the policy optimized in Equation (9) from the policy used to collect data and in the regret statement.

The KL orientation is KL(fitted actor$\|$teacher). Original MPO fits KL(teacher$\|$actor), with an additional policy constraint (Abdolmaleki et al., 2018). We call the shared baseline update projected mirror descent (PMD). The nonconvex recurrent class precludes assuming an exact Bregman projection. The influence loss is a local surrogate over all earlier inputs, not the certified tail $T_h(m)$; it can suppress a useful cue. Section D.2 specifies decision-stage scaling and cue-versus-nuisance audits. Regret uses distortion after exploration mixing; Lemma 14 gives the correction from fitted distortion.

6 Conditional Online Regret Decomposition

The result is a conditional population decomposition on the history-state MDP induced by the POMDP. Write

$$Q_h^\pi(H_h, a) = \mathbb{E}^\pi\left[\sum_{t=h}^{H} R_t \,\middle|\, H_h, A_h = a\right],$$
$$V_h^\pi(H_h) = \sum_a \pi_h(a \mid H_h) Q_h^\pi(H_h, a), \qquad A_h^\pi = Q_h^\pi - V_h^\pi.$$

At epoch k, policy π_k is deployed for m_k episodes, so

$$N_K = \sum_{k=1}^{K} m_k, \qquad \mathrm{Regret}_{N_K}(\pi^\star) = \sum_{k=1}^{K} m_k\{V^{\pi^\star} - V^{\pi_k}\}.$$

Let $d_h^\pi = \mathcal{L}_\pi(H_h)$ and define the comparator potential

$$\Psi_k = \sum_{h=1}^{H} \mathbb{E}_{H_h \sim d_h^{\pi^\star}} \mathrm{KL}\{\pi_h^\star(\cdot \mid H_h) \| \pi_{k,h}(\cdot \mid H_h)\}.$$

For the teacher in Equation (8), define

$$\xi_k = \sum_{h=1}^{H} \sup_{H_h} \max_{a \in \mathcal{A}} \left[\log \frac{\widetilde{\pi}_{k+1,h}(a \mid H_h)}{\pi_{k+1,h}(a \mid H_h)}\right]_+. \tag{11}$$

Let H_{dec} count stages at which some admissible history has more than one action. For the fitted base actor, define $\bar{\xi}_k$ by replacing $\pi_{k+1,h}$ in Equation (11) with $\bar{\pi}_{k+1,h}$. Because $\pi_{k+1,h}(a \mid H_h) \geq (1 - \gamma_{k+1})\bar{\pi}_{k+1,h}(a \mid H_h)$,

$$\xi_k \leq \bar{\xi}_k + H_{\mathrm{dec}} \log \frac{1}{1 - \gamma_{k+1}}. \tag{12}$$

Forced-action stages contribute zero. Thus the pre-mixture projection objective and the deployed-policy distortion differ by a known exploration term. This deterministic bridge does not remove the need to relate sampled $\widehat{\bar{\xi}}_k$ to population $\bar{\xi}_k$.

If the learner policies have full action support and $\xi_k < \infty$, then for every comparator with finite initial potential,

$$\sum_h \mathbb{E}_{d_h^{\pi^\star}}\left[\mathrm{KL}(\pi_h^\star \| \pi_{k+1,h}) - \mathrm{KL}(\pi_h^\star \| \widetilde{\pi}_{k+1,h})\right] \leq \xi_k. \tag{13}$$

This functional discrepancy controls potential transfer for every comparator.

Define the weighted critic discrepancy

$$\epsilon_{Q,k} = m_k \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^\star}} \left|\left\langle A_h^{\pi_k} - \hat{A}_{k,h}, \pi_h^\star - \pi_{k,h} \right\rangle\right|. \tag{14}$$

A uniform advantage error $\|\hat{A}_{k,h} - A_h^{\pi_k}\|_\infty \leq \bar{\epsilon}_{Q,k}$ gives $\epsilon_{Q,k} \leq 2H m_k \bar{\epsilon}_{Q,k}$.

Theorem 10 (Batch-aligned projected mirror descent)**.** *Suppose $\Psi_1 < \infty$, every learner policy has full action support, $\xi_k < \infty$, and $\hat{A}_{k,h}$ is centered under $\pi_{k,h}$. If*

$$\mathrm{osc}_a\{\hat{A}_{k,h}(H_h, a)\} \leq V_h^{\mathrm{osc}}$$

uniformly in k, H_h, then for constant $\eta > 0$,

$$\mathrm{Regret}_{N_K}(\pi^\star) \leq \frac{\Psi_1 + \sum_{k=1}^{K} \xi_k}{\eta} + \frac{\eta}{8}\left(\sum_{k=1}^{K} m_k^2\right)\left(\sum_{h=1}^{H} (V_h^{\mathrm{osc}})^2\right) + \sum_{k=1}^{K} \epsilon_{Q,k}. \tag{15}$$

Clipping permits $V_h^{\mathrm{osc}} = H - h + 1$, giving $\sum_h (V_h^{\mathrm{osc}})^2 = H(H+1)(2H+1)/6$. Forced-action stages instead have zero oscillation. For E1–E3 only the terminal term remains, so this sum can be replaced by one.

This is a population decomposition, not an unconditional learning rate. Lemma 15, Proposition 16, and Corollary 17 connect Monte Carlo fitting to the critic term under explicit coverage and optimization conditions. Lemma 13 links deployed-policy population KL to potential transfer; Lemma 18 controls the sampled-to-uniform gap when a genuine history cover is available. A finite distortion or small training loss alone does not give sublinear regret. Sensitivity penalties enter this analysis only through separately justified control of fitting errors or comparator approximation.

Representation error enters separately through the best-in-class comparator term in Proposition 19; it is not inserted into the mirror-descent telescope. The complementary delayed-cue result (Theorem 20) shows that if $\Gamma_{H,1}$ makes the two terminal action laws nearly indistinguishable, every episode loses nearly one-half unit of value. This construction diagnoses excessive regularization; it lies outside the contractive-filter class and is not a matching lower bound.

7 Empirical Evaluation

We evaluate delayed recall (E1), conditional recall (E2), a filter-stable chain (E3), and active sensing (E4); see Section E. Cumulative pseudo-regret $\sum_k m_k(V^\star - V^{\pi_k})$ is primary; final value is secondary. The archived runs verify 100 updates of 50 episodes ($N_K = 5{,}000$). Baselines and complete ablations use 20 paired seeds; four E4 ablations have only five complete records, marked in the tables. Intervals are 95% seed-bootstrap intervals. The executed protocol uses fixed hyperparameters, decision-stage losses, fitted-Q targets, and information losses on both networks; Section D distinguishes it from the reference formulation. No matched validation search is documented.

The *–PMD* labels describe the actor-to-teacher KL update in the archived *–MPO* runs. Model-free controls vary the encoder while sharing the critic architecture and teacher. GAMPO-KF uses the Bayes operator; GAMPO-P-R remains unevaluated. Table 1 compares regret under this protocol.

Table 1: Cumulative pseudo-regret (mean and 95% seed-bootstrap interval; lower is better). The protocol specifies matched budgets; run-level verification remains outstanding. GAMPO-KF is model-informed.

Method	E1: delayed recall	E2: conditional recall	E3: filter stable	E4: active sensing
GAMPO-KF (ours)	792.8 [763.6, 828.3]	1987.0 [1916.4, 2060.4]	688.1 [678.1, 698.1]	2405.3 [2387.0, 2417.2]
GAMPO-P (ours)	1998.4 [1996.2, 2000.6]	**2479.2** [2467.6, 2489.0]	1159.5 [1142.7, 1175.5]	1996.0 [1963.8, 2027.5]
LSTM–PMD	1999.0 [1996.9, 2001.1]	2488.4 [2477.3, 2496.6]	996.0 [982.1, 1011.0]	1512.5 [1462.5, 1566.6]
GRU–PMD	1999.6 [1996.8, 2002.3]	2480.9 [2456.8, 2497.3]	**925.9** [909.9, 941.3]	1519.8 [1481.3, 1558.3]
Transformer–PMD	**1999.3** [1996.9, 2001.5]	2499.2 [2497.3, 2501.0]	1345.3 [1309.9, 1383.6]	**1241.7** [1212.7, 1271.7]
S4D–PMD	2000.5 [1998.8, 2002.0]	2495.7 [2491.9, 2498.9]	1270.6 [1247.1, 1291.4]	1737.0 [1710.7, 1761.5]
Windowed MLP–PMD	1998.7 [1996.6, 2000.9]	2495.9 [2492.1, 2499.2]	1552.3 [1534.7, 1569.2]	2005.5 [1984.0, 2026.4]
Reactive MLP–PMD	1999.8 [1997.5, 2002.2]	2500.5 [2498.6, 2502.4]	1489.8 [1482.4, 1497.1]	2449.7 [2441.7, 2457.6]

GAMPO-P finishes near chance on E1 (0.499 versus optimum 0.900) and E2 (0.520 versus 1.000), and trails GRU–PMD on E3 (0.730 versus 0.794) and Transformer–PMD on E4 (0.785 versus 0.936). GAMPO-KF leads E1–E3, but uses the Bayes operator. These outcomes do not establish successful model-free long-memory control.

Table 2 isolates the regularization terms. On E4, omission changes regret by -25.3 (95% CI $[-55.2, 2.9]$); on E3 it improves regret by 200.1 [189.3, 211.7]. The 20.1% reduction relative to the forget-gate proxy on E4 therefore establishes only that the proxy is worse under this configuration. On E1 both variants remain at chance, with no reliable regret difference.

Table 2: Paired change in cumulative pseudo-regret, $\Delta = \mathrm{Regret}_{\mathrm{ablated}} - \mathrm{Regret}_{\mathrm{full}}$ (mean and 95% bootstrap interval). Positive values favor the full method. Entries use 20 paired seeds, except $^{(5)}$ uses five and the reference restricted to the same seeds.

Ablation	E1 Δ	E2 Δ	E3 Δ	E4 Δ
Full GAMPO-P (ours)	—	—	—	—
	0.5	-4.2	-200.1	-25.3
$\lambda_\Gamma = 0$	$[-2.7,\ 3.5]$	$[-26.9,\ 15.5]$	$[-211.7,\ -189.3]$	$[-55.2,\ 2.9]$
	0.2	-0.5	-195.2	503.6
Forget-gate proxy	$[-0.5,\ 0.7]$	$[-15.8,\ 13.2]$	$[-210.6,\ -178.4]$	$[470.9,\ 538.0]$
	1.2	13.5	1.1	-127.0
$\lambda_K = \lambda_Q = 0$	$[-1.9,\ 4.3]$	$[5.3,\ 22.5]$	$[-12.2,\ 14.0]$	$[-185.4,\ -84.8]^{(5)}$
	0.5	12.6	46.3	-69.8
$\lambda_B = 0$	$[-3.5,\ 4.1]$	$[1.5,\ 24.8]$	$[35.5,\ 56.5]$	$[-118.5,\ -21.1]^{(5)}$
	0.3	6.6	1.8	-90.0
$\lambda_D = 0$	$[-0.5,\ 1.2]$	$[-0.5,\ 14.5]$	$[-7.1,\ 10.6]$	$[-131.5,\ -31.4]^{(5)}$
	0.6	9.2	-163.6	-483.5
LSTM–PMD	$[-2.6,\ 3.6]$	$[-7.7,\ 25.5]$	$[-181.1,\ -145.9]$	$[-541.4,\ -421.6]$
Full GAMPO-KF (reference)	—	—	—	—
	952.6	410.1	444.6	-235.9
KF without consistency	[782.2, 1107.4]	[317.8, 493.5]	[297.2, 619.0]	$[-439.4,\ -22.3]^{(5)}$

Learning curves and diagnostics (Section A.4) clarify these failures. E2's saved actors attain long-/short-context accuracies 0.500/0.537. Shortening E1's delay to eight lets both variants learn; removing distractors alone at delay 64 does not. All seven follow-up regret intervals include zero. On E4, an exact-belief actor improves final value from 0.914 to 0.956 but increases regret from 1512.2 to 1807.4; exact beliefs in both networks also fail to improve regret. These controls use five new paired seeds and the same training budget, exposing remaining learning inefficiency even without learned information states. A separate fast-refresh control ($H = 8, \rho = \zeta = 0, d_N = 64$) confirms a narrower benefit: with $\lambda_\Gamma = 1$, regret falls by 300.8 [272.2, 327.4] over omission across 20 new paired seeds (Section A.5); past information is unnecessary here.

8 Discussion and Conclusion

Full joint-state Jacobians expose memory paths omitted by forget gates. Uniform envelopes bound finite perturbations, while calibration and filter stability connect them to control. A terminal-decision refinement removes unnecessary horizon factors, and the online analysis separates critic, projection, and representation errors under explicit conditions.

The four controlled tasks support full-path sensitivity as a diagnostic, but do not establish a general optimization advantage for GAMPO-P. The independently confirmed fast-refresh benefit concerns suppression of unnecessary history, not recovery of long-term memory. The separate terminal-state head in GAMPO-P-R remains unevaluated; the archived implementation already predicts terminal rewards in extended blocks. Priority checks are objective and normalization ablations, cue-versus-distractor audits, and an independently validated recurrent RL baseline (Section F). Broader evaluation requires POPGym (Morad et al., 2023), GTrXL and Mamba baselines (Parisotto et al., 2020; Gu & Dao, 2023), and measured compute costs.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1ANxQW0b.

Tankut Can, Kamesh Krishnamurthy, and David J. Schwab. Gating creates slow modes and controls phase-space complexity in GRUs and LSTMs. In *Proceedings of the First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 476–511. PMLR, 2020. URL https://proceedings.mlr.press/v107/can20a.html.

Semih Cayci and Atilla Eryilmaz. Recurrent natural policy gradient for POMDPs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=6G01e0vgIf.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 349–362, 2023. doi: 10.1145/3564246.3585099. URL https://doi.org/10.1145/3564246.3585099.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Re. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35971–35983, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e9a32fade47b906de908431991440f7c-Abstract-Conference.html.

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://cdn.aaai.org/ocs/11673/11673-51288-1-PB.pdf.

Ahmed Hefny, Zita Marinho, Wen Sun, Siddhartha Srinivasa, and Geoffrey Gordon. Recurrent predictive state policy networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1949–1958. PMLR, 2018. URL https://proceedings.mlr.press/v80/hefny18a.html.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

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.

Ali Devran Kara and Serdar Yüksel. Near optimality of finite memory feedback policies in partially observed markov decision processes. *Journal of Machine Learning Research*, 23(11):1–46, 2022. URL https://www.jmlr.org/papers/v23/20-1152.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Rodney Lafuente-Mercado, Daniela Rus, and T. Konstantin Rusch. Quantifying memory use in reinforcement learning with temporal range. *arXiv preprint arXiv:2512.06204*, 2025. URL https://arxiv.org/abs/2512.06204.

Lorenzo Livi. Learnability window in gated recurrent neural networks. *Physical Review E*, 114(3):035301, 2026. doi: 10.1103/843n-yshj. URL https://arxiv.org/abs/2512.05790.

Curtis McDonald and Serdar Yüksel. Exponential filter stability via dobrushin's coefficient. *Electronic Communications in Probability*, 25(53):1–12, 2020. doi: 10.1214/20-ECP333.

John Miller and Moritz Hardt. Stable recurrent models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hygxb2CqKm.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=chDrutUTs0K.

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, pp. 7487–7498. PMLR, 2020. URL https://proceedings.mlr.press/v119/parisotto20a.html.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16(6):155–186, 2015. URL https://www.jmlr.org/papers/v16/rakhlin15a.html.

Uri Sherman, Tomer Koren, and Yishay Mansour. Convergence of policy mirror descent beyond compatible function approximation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54825–54863. PMLR, 2025. URL https://proceedings.mlr.press/v267/sherman25a.html.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022. URL https://www.jmlr.org/papers/v23/20-1165.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Arun Venkatraman, Nicholas Rhinehart, Wen Sun, Lerrel Pinto, Martial Hebert, Byron Boots, Kris M. Kitani, and J. Andrew Bagnell. Predictive-state decoders: Encoding the future into recurrent networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1709.08520.

Andrew Wang, Andrew C. Li, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. Learning belief representations for partially observable deep RL. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35970–35988. PMLR, 2023. URL https://proceedings.mlr.press/v202/wang23p.html.

SUPPLEMENTARY MATERIAL

A ADDITIONAL EXPERIMENTAL RESULTS

A.1 COMPLETE CONTROLLED-TASK RESULTS

Tables 3 and 4 report cumulative regret and final value, respectively. Most entries use 20 archived seeds (0–19). Four E4 ablations have only five complete archived runs (seeds 0–4), marked $^{(5)}$; paired contrasts use the same five reference seeds. An isolated additional sensitivity-ablation record is not treated as a 20-seed cohort. GAMPO-KF reaches final values 0.876 and 0.854 on E1 and E2, while GAMPO-P remains near chance. On E4, removing the influence penalty lowers regret slightly but reduces final value from 0.785 to 0.744. An endpoint gain alone therefore does not establish faster learning.

Table 3: Complete cumulative pseudo-regret (lower is better). Means and 95% seed-bootstrap intervals use 20 seeds, except $^{(5)}$ uses five. GAMPO-KF is model-informed; all GAMPO ablations are variants of our methods.

Method or variant	E1	E2	E3	E4
GAMPO-KF (ours)	792.8 [763.6, 828.3]	1987.0 [1916.4, 2060.4]	688.1 [678.1, 698.1]	2405.3 [2387.0, 2417.2]
GAMPO-P (ours)	1998.4 [1996.2, 2000.6]	2479.2 [2467.6, 2489.0]	1159.5 [1142.7, 1175.5]	1996.0 [1963.8, 2027.5]
LSTM–PMD	1999.0 [1996.9, 2001.1]	2488.4 [2477.3, 2496.6]	996.0 [982.1, 1011.0]	1512.5 [1462.5, 1566.6]
GRU–PMD	1999.6 [1996.8, 2002.3]	2480.9 [2456.8, 2497.3]	925.9 [909.9, 941.3]	1519.8 [1481.3, 1558.3]
Transformer–PMD	1999.3 [1996.9, 2001.5]	2499.2 [2497.3, 2501.0]	1345.3 [1309.9, 1383.6]	1241.7 [1212.7, 1271.7]
S4D–PMD	2000.5 [1998.8, 2002.0]	2495.7 [2491.9, 2498.9]	1270.6 [1247.1, 1291.4]	1737.0 [1710.7, 1761.5]
Windowed MLP–PMD	1998.7 [1996.6, 2000.9]	2495.9 [2492.1, 2499.2]	1552.3 [1534.7, 1569.2]	2005.5 [1984.0, 2026.4]
Reactive MLP–PMD	1999.8 [1997.5, 2002.2]	2500.5 [2498.6, 2502.4]	1489.8 [1482.4, 1497.1]	2449.7 [2441.7, 2457.6]
GAMPO-P, $\lambda_\Gamma = 0$	1998.9 [1996.5, 2001.0]	2475.1 [2452.0, 2493.5]	959.4 [943.5, 975.7]	1970.7 [1944.4, 1993.6]
GAMPO-P, forget-gate proxy	1998.5 [1996.5, 2000.6]	2478.7 [2458.7, 2495.0]	964.3 [952.5, 978.1]	2499.6 [2494.3, 2504.6]
GAMPO-P, $\lambda_K = \lambda_Q = 0$	1999.5 [1997.2, 2001.7]	2492.7 [2486.3, 2498.6]	1160.7 [1146.2, 1175.8]	1879.8 [1849.7, 1908.2]$^{(5)}$
GAMPO-P, $\lambda_B = 0$	1998.9 [1996.1, 2001.6]	2491.8 [2486.9, 2496.6]	1205.8 [1195.3, 1216.1]	1937.0 [1889.6, 1991.8]$^{(5)}$
GAMPO-P, $\lambda_D = 0$	1998.7 [1996.5, 2000.8]	2485.9 [2474.4, 2494.6]	1161.4 [1144.1, 1179.0]	1916.8 [1867.9, 1970.2]$^{(5)}$
GAMPO-KF without consistency	1745.4 [1573.0, 1901.0]	2397.2 [2346.7, 2441.9]	1132.7 [982.7, 1308.5]	2182.9 [1972.5, 2403.4]$^{(5)}$

A.2 REGULARIZATION TRADEOFFS AND ACTIVE SENSING

Removing the complete GAMPO-P package lowers regret by 163.6 on E3 and 483.5 on E4. E3's influence ablation is consistent with suppression of useful sensitivity alongside nuisance dependence. Several E4 ablations suggest competing objectives but do not identify a unique mechanism. GAMPO-KF has E4 regret 2405.3 and final value 0.590. In the five matched seeds available for its consistency ablation, removing consistency lowers regret by 235.9. The exact-belief controls in Section A.4 test whether directly supplying sufficient information resolves this failure.

Table 4: Complete held-out final-policy value (higher is better). Means and 95% seed-bootstrap intervals use 20 seeds, except [5] uses five. GAMPO-KF is model-informed; all GAMPO ablations are variants of our methods.

Method or variant	E1	E2	E3	E4
GAMPO-KF (ours)	0.876 [0.874, 0.879]	0.854 [0.800, 0.905]	0.843 [0.841, 0.846]	0.590 [0.565, 0.626]
GAMPO-P (ours)	0.499 [0.497, 0.502]	0.520 [0.511, 0.529]	0.730 [0.725, 0.735]	0.785 [0.760, 0.808]
LSTM–PMD	0.499 [0.497, 0.500]	0.508 [0.502, 0.515]	0.777 [0.773, 0.781]	0.894 [0.882, 0.905]
GRU–PMD	0.500 [0.498, 0.501]	0.515 [0.504, 0.532]	0.794 [0.791, 0.797]	0.905 [0.898, 0.913]
Transformer–PMD	0.501 [0.498, 0.503]	0.500 [0.498, 0.503]	0.718 [0.705, 0.729]	0.936 [0.931, 0.941]
S4D–PMD	0.499 [0.497, 0.501]	0.501 [0.499, 0.504]	0.726 [0.719, 0.734]	0.885 [0.871, 0.899]
Windowed MLP–PMD	0.499 [0.497, 0.501]	0.503 [0.501, 0.506]	0.641 [0.631, 0.650]	0.792 [0.784, 0.801]
Reactive MLP–PMD	0.500 [0.498, 0.502]	0.502 [0.500, 0.504]	0.617 [0.614, 0.619]	0.586 [0.576, 0.595]
GAMPO-P, $\lambda_\Gamma = 0$	0.500 [0.497, 0.502]	0.519 [0.505, 0.536]	0.787 [0.783, 0.791]	0.744 [0.721, 0.769]
GAMPO-P, forget-gate proxy	0.499 [0.497, 0.502]	0.520 [0.505, 0.538]	0.784 [0.781, 0.787]	0.540 [0.537, 0.543]
GAMPO-P, $\lambda_K = \lambda_Q = 0$	0.499 [0.498, 0.501]	0.508 [0.503, 0.515]	0.728 [0.723, 0.734]	0.864 [0.847, 0.878][5]
GAMPO-P, $\lambda_B = 0$	0.499 [0.497, 0.502]	0.508 [0.504, 0.513]	0.712 [0.707, 0.718]	0.781 [0.759, 0.802][5]
GAMPO-P, $\lambda_D = 0$	0.498 [0.497, 0.500]	0.515 [0.508, 0.524]	0.728 [0.721, 0.735]	0.846 [0.819, 0.880][5]
GAMPO-KF without consistency	0.626 [0.553, 0.702]	0.618 [0.582, 0.656]	0.732 [0.691, 0.765]	0.642 [0.565, 0.718][5]

Table 5: E3 filtering-to-control bound audit at the locked final GAMPO-KF policy π_{K+1}. Exact-tree rows enumerate all 3^H histories of $(H, \rho, \zeta, d_N) = (12, 0.90, 0.75, 0)$ and weight them by their exact probabilities; Monte Carlo rows use 10,000 held-out episodes of the primary setting $(32, 0.90, 0.75, 16)$. TV $= |\hat{b}_h - b_h|$ is pooled over histories and stages; "defect" is the held-out maximum of the one-step defect δ_h in Equation (3) and "filter RHS" the held-out maximum of the pathwise right-hand side of Theorem 5, so the row reads $\max \text{TV} \le$ filter RHS. "Violation" counts seeds in which the observed error exceeds the computed bound at any (history, stage) beyond the tolerance 10^{-6}. The value gap is $|V^{\hat{\pi}} - V^{\pi^b}|$ for the same terminal head applied to the decoded and to the exact belief. The value-bound column reports the raw right-hand side of Equation (6) and, after the slash, its minimum with the known unit range of total return. Here e_s is the held-out maximum at stage s and L_π the Lipschitz constant of the head. Entries are means with 95% seed-bootstrap intervals over 20 test seeds.

Audit	mean TV	p95 TV	max TV	defect	filter RHS	violation	value gap	value bound raw / clipped
Exact tree	0.006 [0.005, 0.008]	0.026 [0.021, 0.033]	0.290 [0.249, 0.331]	0.221 [0.194, 0.249]	0.337 [0.298, 0.379]	0/20	0.0014 [0.0011, 0.0018]	23.5 [21.4, 25.7] / 1.0
Monte Carlo	0.017 [0.016, 0.018]	0.076 [0.071, 0.081]	0.344 [0.318, 0.371]	0.209 [0.193, 0.227]	0.359 [0.331, 0.386]	0/20	0.0035 [0.0031, 0.0039]	251.1 [240.5, 262.7] / 1.0

A.3 MECHANISTIC AND CRITIC AUDITS

Table 5 compares filtering and control errors in E3, where the Bayes contraction coefficient is $\rho_B = 0.9$. Mean belief TV is 0.006 on the exact tree and 0.017 under Monte Carlo, while the maximum error is much larger (0.290 and 0.344). No pathwise filtering-bound violation is reported in either audit. The maxima of the computed filtering right-hand side (0.337 and 0.359) are close to the corresponding error maxima. These aggregate statistics do not identify whether isolated defects or accumulated defects dominate individual trajectories.

Applying the same head to exact rather than decoded beliefs changes value by only 0.0014 and 0.0035, respectively. The generic value expressions (23.5 and 251.1) are vacuous for this task: they sum worst-history errors over all stages, whereas E3 has only a terminal decision and total return at most one. The range-clipped bound is therefore the trivial value 1.0 in both rows. The exact-tree calculation covers its stated finite setting. The Monte Carlo row substitutes sampled maxima for the suprema in Theorem 8, so it is a diagnostic rather than a uniform certificate. Neither the small observed value gap nor the absence of sampled violations establishes the hypotheses outside the audited histories. Corollary 9 instead uses terminal-stage belief error and a verified head Lipschitz constant. The pooled summaries in this table do not supply the per-seed terminal errors needed to compute that bound, so no replacement numerical certificate is inferred from them.

Table 6: Recomputed E0 audit for the analytic cell in Section E.1. The 80 combinations of $f \in \{0.05, 0.2, 0.5, 0.8\}$, $u \in \{0, 0.5, 0.9, 1.1\}$, and $D \in \{2, 4, 8, 16, 32\}$ compare the effect $I = I^Z_{D+1,1}$ for $X_1 = \pm 0.1$ with diagnostics on the positive-input trajectory. The Temporal Range contribution is $D^{-1} \sum_{s=2}^{D+1} \|\partial Z_s / \partial X_1\|$. R^2 regresses $\log I$ on the log diagnostic; ratios are diagnostic$/I$. Violation tests are $I -$ diagnostic $> 10^{-6}$ (absolute) and $1 -$ diagnostic$/I > 10^{-6}$ (relative). All quantities use double precision; no RL policy is fitted.

Diagnostic	R^2	median ratio	p95 ratio	absolute	relative
Jacobian product	0.972	0.964	0.993	0.850	1.000
Norm envelope	0.946	1.882	2.3×10^4	0.300	0.300
Forget-only	0.293	0.007	0.989	0.875	1.000
Temporal contribution	0.651	0.993	2.8×10^8	0.500	0.500

The earlier single-column convention is replaced by explicit absolute and relative tests. Absolute violation counts are 68, 24, 70, and 40 out of 80; relative counts are 80, 24, 80, and 40. At $f = 0.05, u = 0, D = 32$, $I^Z_{D+1,1} \simeq 3.04 \times 10^{-42}$; no discrepancy from a nonnegative diagnostic can exceed the absolute tolerance.

Table 6 isolates input influence without training a policy. The endpoint full-path Jacobian has the highest log-scale agreement with finite interventions ($R^2 = 0.972$, median ratio 0.964), but underestimates the finite effect in every grid cell. An endpoint derivative need not bound a finite perturbation across an interval. The sample-local norm envelope dominates the Jacobian product by submultiplicativity, yet still underestimates 30% of finite interventions and has a 95th-percentile ratio of 2.3×10^4. It uses pointwise norms, not the domain-wide suprema required by Proposition 1.

The forget-only proxy has much weaker agreement ($R^2 = 0.293$, median ratio 0.007), consistent with omitting the hidden-state-to-candidate route when $u > 0$. The Temporal Range contribution averages intermediate-lag influence; its aggregation targets an average over intermediate lags, not an upper bound on a terminal intervention. Its violation rate is a scale comparison, not a failure of a claimed Temporal Range guarantee. These results motivate full-path diagnostics while preserving the distinction between local derivatives, finite effects, and global certificates.

Table 7 reports local quantities at the final checkpoint used in Table 4. In every row, the local norm-product estimate of effective memory reaches $H - 1$ after subtracting one. This saturation does not establish that the policy uses every past input; it indicates that this envelope fails to certify a shorter window at the chosen threshold. The E3 belief-error 95th percentile of 0.076 agrees with Table 5, but a quantile does not verify the uniform representative-history calibration assumed by Corollary 6. Temporal Range is shorter on E3 (2.8) than on E4 (15.4), consistent with their different information demands; this remains a descriptive comparison.

Table 7: Mechanistic diagnostics at the locked final checkpoint π_{K+1}. Held-out quantities use 1,000 fresh episodes and the Jacobian quantities the first 256 of them, with ten power iterations. $\hat{m}_{\mathrm{eff}}(H;\epsilon) - 1$ is Definition 2 at the terminal stage with the sample-local envelope $\hat{\Gamma}_{H,s} = \hat{J}^{\mathrm{env}}_{H,s}$, the exact input-domain diameter D_X, and the prespecified threshold $\epsilon = 0.1$; belief TV is the 95th percentile of $\mathrm{TV}(\hat{b}_h, b_h)$ and is defined only for the model-informed decoder; Temporal Range is $\hat{\rho}_H$ of the actor logits at the admissible-decision stages ($L = \mathrm{mean}$); $\hat{K}^{\mathrm{loc}}_A$ is the held-out maximum of $\|\partial \ell_h / \partial \vartheta_A\|_{\mathrm{op}}$; and $\sum_k \widehat{\widehat{\xi}}_k$ accumulates the pre-mixture distortion in Equation (28) with the exact maximum over the sampled batch after every actor update. The effective-memory entry is a held-out maximum; belief TV is a percentile. The remaining entries aggregate the stated per-seed diagnostics as means with 95% seed-bootstrap intervals over 20 test seeds. Local quantities are diagnostics, not global certificates.

Env.	Variant	$\hat{m}_{\mathrm{eff}}(H;\epsilon) - 1$	belief TV (p95)	Temporal Range	$\hat{K}^{\mathrm{loc}}_A$	$\sum_k \widehat{\widehat{\xi}}_k$
E1	GAMPO-P	64	—	5.2 [3.1, 8.5]	5.3 [2.0, 11.7]	5.2 [4.5, 6.2]
E1	LSTM–PMD	64	—	8.0 [6.4, 10.1]	2.8 [2.4, 3.3]	3.6 [3.1, 4.1]
E2	GAMPO-P	64	—	5.9 [4.8, 7.0]	7.5 [5.3, 10.0]	9.3 [8.1, 10.4]
E3	GAMPO-KF	31	0.076 [0.071, 0.081]	2.8 [2.7, 2.9]	35.3 [30.5, 41.6]	17.4 [16.8, 17.9]
E4	GAMPO-P	31	—	15.4 [15.1, 15.7]	37.1 [33.4, 41.0]	50.0 [49.1, 50.8]

E4 has the largest reported actor sensitivity and accumulated sampled pre-mixture projection distortion. However, $\sum_k \widehat{\widehat{\xi}}_k$ is not the population $\sum_k \xi_k$ in Theorem 10; finiteness on sampled histories does not establish a finite uniform bound. For illustration, at $\eta = 0.3/50 = 0.006$, dividing the reported sampled sums by η gives approximately 6.0×10^2 for E1 LSTM–PMD and 8.3×10^3 for E4 GAMPO-P. These are empirical pre-mixture scales, not certified regret contributions; Equation (12) supplies a separate deterministic exploration correction. Finally, both E1 variants remain near chance, and their Temporal Range intervals overlap. Their comparison does not demonstrate successful long-memory control.

Table 8: Held-out critic audit of the GAMPO-P critic using an independent continuation bank. For each of 256 held-out histories, every admissible-decision stage h and every action a, the latent is resampled from the exact posterior b_h, action a is taken at h, and π_{K+1} is followed afterwards for 64 independent continuations; the error is $\hat{q}_h(H_h, a)$, clipped as in Equation (23), minus the continuation mean. Monte Carlo continuation standard error is reported separately from critic error. $\hat{K}^{Q,\mathrm{loc}}$ is the held-out maximum of $\max_a \|\nabla_{\vartheta_Q} q_h(H_h, a)\|_2$. The last two columns regress $|\mathrm{error}|$ on the logarithm of the local envelope $\hat{K}^{Q,\mathrm{env,loc}}$ of Section D.3, computed once with full-Jacobian transport $\prod_j \|\hat{J}^{ZZ}_j\|_{\mathrm{op}}$ and once with forget-only transport $\prod_j \|D(F_j)\|_{\mathrm{op}}$, pooling (history, stage) pairs across seeds with 95% intervals from seed-clustered standard errors. Entries are means with 95% seed-bootstrap intervals over 20 test seeds.

Environment	RMSE	bias	MC SE	$\hat{K}^{Q,\mathrm{loc}}$	full-Jacobian coef.	forget-only coef.
E1	0.432 [0.378, 0.473]	+0.005 [-0.040, +0.052]	0.037	86.2 [51.9, 130.4]	-0.0004 [-0.0077, 0.0069]	0.0825 [0.0028, 0.1621]
E2	0.414 [0.347, 0.470]	-0.028 [-0.062, +0.004]	0.000	62.9 [39.8, 87.5]	0.0250 [0.0133, 0.0367]	0.2179 [0.1869, 0.2488]
E3	0.251 [0.229, 0.275]	+0.014 [-0.008, +0.040]	0.032	9.5 [8.0, 11.2]	0.0111 [0.0005, 0.0216]	0.0997 [0.0758, 0.1235]
E4	0.132 [0.123, 0.141]	-0.076 [-0.091, -0.061]	0.050	23.1 [18.8, 28.1]	-0.0033 [-0.0052, -0.0015]	-0.0124 [-0.0223, -0.0026]

Table 8 reports final-policy critic diagnostics relevant to, but distinct from, $\epsilon_{Q,k}$ in Theorem 10. That term uses comparator-weighted advantage error at every policy epoch; a held-out final-policy RMSE does not estimate it directly. Continuation standard errors are smaller than the reported RMSEs, although on E4 the standard error of 0.050 is substantial relative to RMSE 0.132. The E2 standard error is reported as 0.000 at the displayed precision. E4 also has a negative mean residual of -0.076; the table does not identify its cause. In particular, it could reflect bootstrapped target propagation

in the executed fitted-Q critic, but this audit cannot isolate that explanation from approximation or optimization error.

Local parameter sensitivities are larger in the two $H = 65$ settings than in the two $H = 32$ settings, but horizon and task differ jointly. The regression slopes are positive on E2 and E3 and negative on E4 for both sensitivity summaries; these associations do not establish a causal relation or a universal error law. Raw slope magnitudes also depend on the spread of each regressor, so they do not alone rank predictive quality. Proposition 11 concerns a global sensitivity certificate, whereas these regressors are local envelopes. This distinction is consistent with the finite-intervention audit: describing which histories a critic fits poorly and bounding retained input influence are different questions.

A.4 Learning curves and targeted failure diagnostics

Figure 2 shows the original 20-seed trajectories, recovered from logged evaluations and matched to final result records. We retain every logged checkpoint (epochs 1, 11, . . ., 91, 100) and the final policy at epoch 101; intermediate logged values have precision 0.001. Bands are pointwise seed-bootstrap intervals, not simultaneous bands.

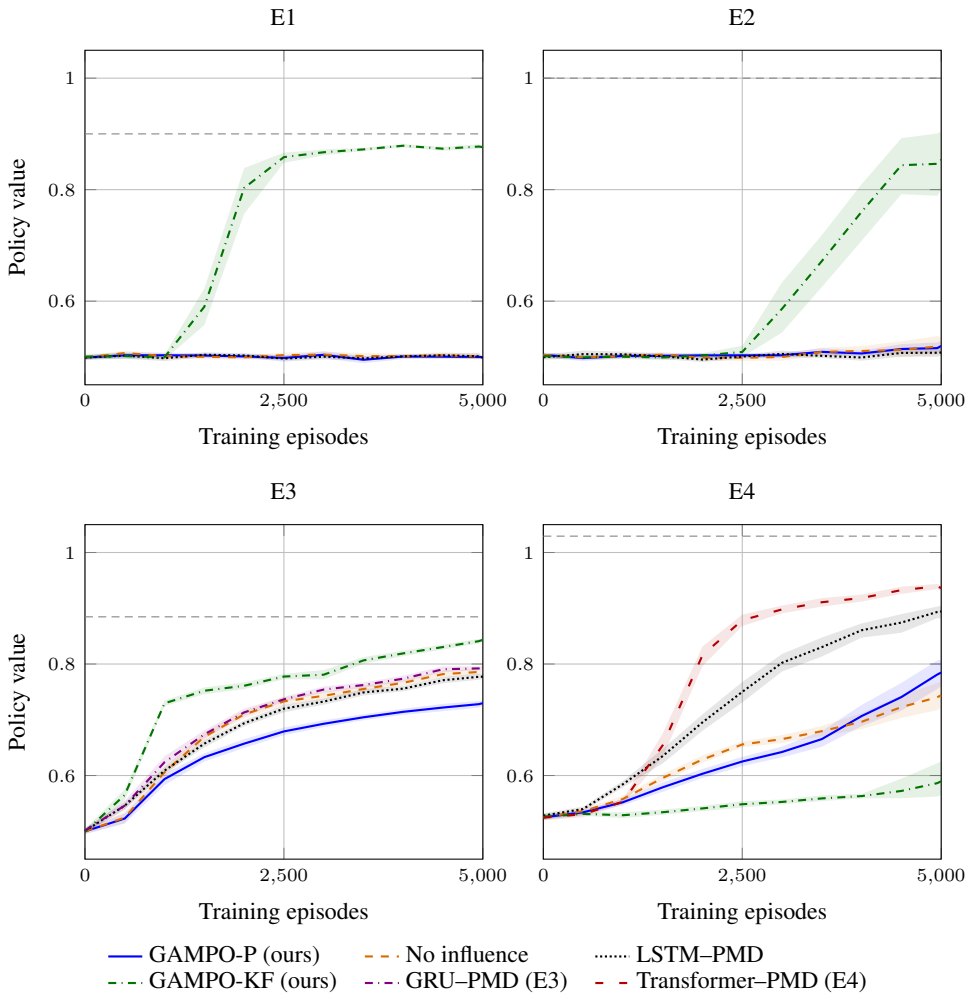


Figure 2: Original learning curves: means and pointwise 95% seed-bootstrap bands over 20 seeds. Gray dashed lines mark optimal value. GAMPO-KF uses model information. E4's endpoint advantage over omission does not imply an advantage in cumulative regret.

We re-evaluate 20 saved GAMPO-P actors from the archived E1 and E2 audit runs with 10,000 fresh episodes per seed. Evaluation integrates terminal correctness against the exact posterior, reducing noise without supplying beliefs to the trained actor. E1 value is 0.4997 (95% CI [0.4992, 0.5001]). E2 long-context accuracy is 0.5005 [0.4997, 0.5013] and short-context accuracy is 0.5372 [0.5222, 0.5541]. These fresh conditional-expectation estimates do not replace the sampled-return entries in Table 4. The failure extends to the four-step short branch, not only the initial cue.

Supplementary runs use seeds 200–204, 100 updates of 50 episodes, and the executed configuration in Section D. Full/omission pairs change only λ_Γ. Hyperparameters and the seven-condition grid are fixed before execution, with no selection on outcomes. Evaluations use 2,000 episodes per epoch and 10,000 for the final policy, integrate terminal correctness, and isolate evaluation random numbers from training. These are controlled follow-ups, not exact seedwise reproductions of the original

execution. Five-seed intervals have limited resolution and the multiple contrasts are exploratory. We report every condition, including null and negative contrasts.

Tables 10 and 11 distinguish overall final value from branch-specific recall. Shortening E1 to $D = 8$ permits both variants to learn with or without 16 distractors. At $D = 16$, learning becomes unstable; at $D = 64$, removing distractors alone leaves both variants at chance. E2 likewise learns both branches at $H = 9$, whereas at $H = 17$ GAMPO-P's long branch remains at chance and its short branch improves. At $H = 65$, even the four-step short branch fails without distractors. This points to learning difficulties on long unrolled trajectories, not only loss of the initial cue, but does not uniquely separate recurrent optimization, critic fitting, and auxiliary-loss competition.

All paired regret intervals in Table 9 contain zero. The favorable mean at $D = 8, d_N = 0$ is too uncertain to establish a benefit over omission. The tiny final-value contrast at $D = 64$ occurs between chance-level policies and does not indicate useful recall. These conditions therefore do not establish a reliable regime in which adding the influence penalty improves the primary endpoint.

Table 9: Recall follow-up: cumulative pseudo-regret, five paired seeds and 5,000 training episodes. Entries give means and 95% seed-bootstrap intervals. Positive $\Delta = \mathrm{Regret}_{\mathrm{omit}} - \mathrm{Regret}_{\mathrm{full}}$ favors GAMPO-P.

Condition	GAMPO-P (ours)	No influence	Paired Δ
E1: $D = 8, d_N = 0$	789.6 [560.7, 1081.1]	962.9 [689.3, 1236.6]	173.4 [−98.0, 521.9]
E1: $D = 8, d_N = 16$	924.4 [829.6, 1056.8]	858.3 [796.3, 924.9]	-66.1 [−171.5, 39.1]
E1: $D = 16, d_N = 0$	1872.7 [1618.8, 1999.9]	1997.0 [1992.3, 1999.7]	124.3 [−7.2, 380.5]
E1: $D = 64, d_N = 0$	1999.5 [1998.9, 2000.2]	1999.7 [1999.2, 2000.2]	0.1 [−0.1, 0.3]
E2: $H = 9, d_N = 0$	909.7 [745.4, 1110.5]	919.3 [839.6, 1007.0]	9.7 [−133.1, 152.4]
E2: $H = 17, d_N = 0$	2294.1 [2143.8, 2445.3]	2290.7 [2130.5, 2450.9]	-3.4 [−243.0, 240.2]
E2: $H = 65, d_N = 0$	2499.9 [2499.3, 2500.6]	2500.1 [2499.8, 2500.4]	0.2 [−0.2, 0.6]

Table 10: Recall follow-up: final-policy value on the same runs, including final exploration. Positive $\Delta = V_{\mathrm{full}} - V_{\mathrm{omit}}$ favors GAMPO-P. Means and 95% seed-bootstrap intervals use five seeds. Optimal value is 0.9 on E1 and 1 on E2.

Condition	GAMPO-P (ours)	No influence	Paired Δ
E1: $D = 8, d_N = 0$	0.876 [0.870, 0.879]	0.873 [0.870, 0.877]	0.002 [−0.003, 0.007]
E1: $D = 8, d_N = 16$	0.867 [0.855, 0.875]	0.868 [0.864, 0.872]	-0.001 [−0.010, 0.008]
E1: $D = 16, d_N = 0$	0.576 [0.500, 0.727]	0.530 [0.500, 0.592]	0.045 [−0.092, 0.227]
E1: $D = 64, d_N = 0$	0.500 [0.499, 0.500]	0.500 [0.499, 0.500]	-0.0002 [−0.0005, −0.0000]
E2: $H = 9, d_N = 0$	0.977 [0.976, 0.977]	0.977 [0.976, 0.977]	0.000 [−0.001, 0.001]
E2: $H = 17, d_N = 0$	0.643 [0.547, 0.739]	0.704 [0.567, 0.840]	-0.061 [−0.275, 0.094]
E2: $H = 65, d_N = 0$	0.500 [0.499, 0.501]	0.500 [0.500, 0.501]	-0.0003 [−0.0007, 0.0000]

For E4, the exact-belief actor receives $(1 - b_h, b_h)$ and the stage index followed by the same two 64-unit tanh layers. Beliefs use only observable history, not the latent state. We compare this actor against LSTM–PMD with the same critic architecture, teacher, optimizer, and budget, and additionally

Table 11: E2 without distractors: final long- and short-context accuracies, evaluated separately. The short-cue delay is fixed at four. Means and 95% seed-bootstrap intervals use five seeds.

Horizon and variant	Long context	Short context
$H = 9$: GAMPO-P	0.977 [0.976, 0.977]	0.977 [0.976, 0.977]
$H = 9$: No influence	0.977 [0.976, 0.977]	0.976 [0.976, 0.977]
$H = 17$: GAMPO-P	0.500 [0.500, 0.501]	0.787 [0.596, 0.979]
$H = 17$: No influence	0.591 [0.500, 0.771]	0.818 [0.629, 0.976]
$H = 65$: GAMPO-P	0.499 [0.498, 0.500]	0.501 [0.500, 0.501]
$H = 65$: No influence	0.500 [0.499, 0.501]	0.500 [0.500, 0.501]

replace both actor and critic representations with exact beliefs. These are model-informed structural controls, not oracle planners: action values and sensing decisions must still be learned.

Replacing the actor representation raises final value but slows learning (Table 12, Figure 3). Relative to LSTM–PMD, final value increases by 0.042 (95% CI [0.027, 0.058]), but regret increases by 295.2 [191.5, 390.7]. Exact beliefs in both networks raise final value to 0.970, but regret still increases by 179.9 [66.6, 293.2]. Final probe fractions remain near one-half. Thus removing learned representations improves the endpoint without resolving learning efficiency or attaining optimal value 1.029. These controls change input dimension and remove recurrent parameters; they do not identify one optimization mechanism, but show that learning inefficiency persists with sufficient information directly available.

Table 12: E4 representation controls: five paired seeds, identical episode and update budgets, means and 95% seed-bootstrap intervals. Probe fraction averages the 31 preterminal stages; terminal accuracy excludes the saving reward.

Representation	Regret	Final value	Probe fraction	Terminal accuracy
Learned actor and critic	1512.2 [1446.1, 1578.4]	0.914 [0.903, 0.925]	0.502 [0.484, 0.518]	0.889 [0.877, 0.901]
Exact actor; learned critic	1807.4 [1747.5, 1869.2]	0.956 [0.949, 0.964]	0.498 [0.488, 0.509]	0.931 [0.924, 0.939]
Exact actor and critic	1692.1 [1641.2, 1739.3]	0.970 [0.969, 0.972]	0.495 [0.491, 0.501]	0.945 [0.944, 0.947]

The 85 unique supplementary runs used NVIDIA A100-SXM4-80GB accelerators in existing HPCC allocations. Recorded per-run elapsed time has median 262 seconds and range 71–1,020 seconds. Runs shared accelerators, so these are not exclusive-device costs or accelerator-hours. Retained records contain all 101 evaluations, subgroup metrics, configuration, seeds, source hashes, and elapsed time. No model checkpoints or best-checkpoint selection are used.

A.5 An independently confirmed fast-refresh regime

We test the hypothesis that penalizing history influence helps when old information rapidly becomes irrelevant and independent distractors are abundant. This study reuses E3 with $H = 8$ and $d_N = 64$, varying $(\rho, \zeta) \in \{(0, 0), (0, 0.5), (0.5, 0.5)\}$ and the actor coefficient $\lambda_\Gamma \in \{0, 0.1, 1\}$. All other network, loss, and optimization settings remain fixed. Each run uses 100 updates and 5,000 training episodes; each pre-update value uses 2,000 independent evaluation episodes, and the final policy uses 10,000. Budgets match episodes and optimizer steps, not elapsed time or FLOPs. The separately weighted terminal head in GAMPO-P-R is not evaluated.

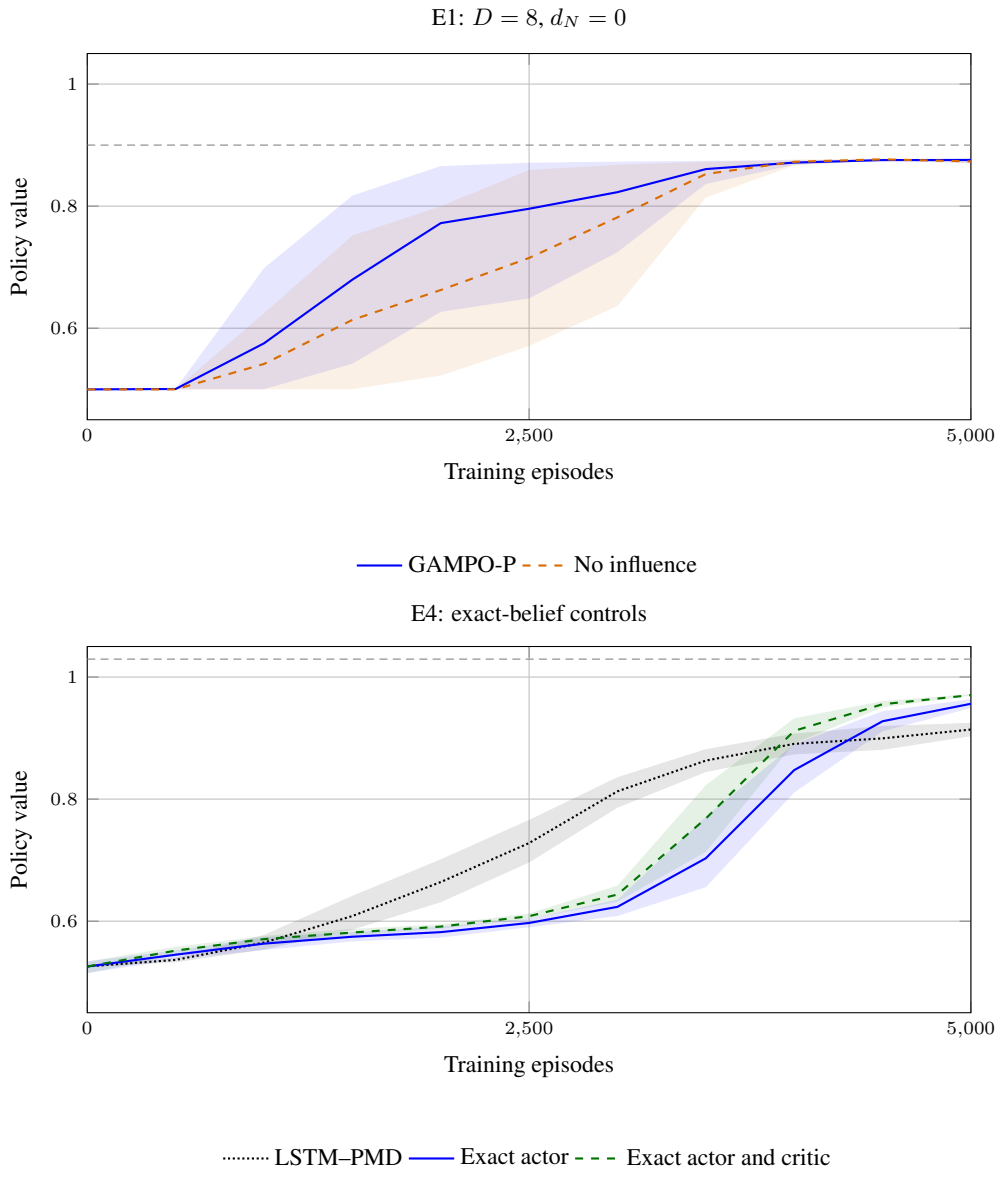


Figure 3: Supplementary learning curves: five paired seeds, means and pointwise 95% bootstrap bands; every tenth update and the final policy are shown. Both variants learn short-delay recall. On E4, exact beliefs improve final value but learn slowly under the same update procedure. Gray lines mark optimal value.

Before pilot training, we fixed five pilot seeds (400–404), the three conditions, and both nonzero coefficients. Among the six contrasts with positive mean regret reduction and final value above 0.55, we select the largest lower endpoint of the paired 95% bootstrap interval (ties broken by condition and method). Table 13 reports all six exploratory contrasts; these intervals do not correct for selection or multiple comparisons. The selected condition and coefficient are then locked for a single confirmation on 20 unused paired seeds (500–519). Success requires a strictly positive lower confidence endpoint for omission-minus-inclusion regret and included-penalty final value above 0.55. No alternative is substituted after confirmation. All intervals use 10,000 seed-level bootstrap resamples with bootstrap seed 6025.

Table 13: Complete exploratory pilot: five paired seeds per contrast. Positive Δ favors including the penalty. These are selection-stage results, not independent confirmation.

ρ	ζ	λ_Γ	Regret Δ (95% CI)
0	0	0.1	136.6 [117.3, 154.8]
0	0	1	301.0 [279.8, 325.4]
0	0.5	0.1	61.8 [32.8, 93.3]
0	0.5	1	150.9 [117.7, 178.3]
0.5	0.5	0.1	1.1 [−10.1, 14.1]
0.5	0.5	1	79.4 [62.4, 96.4]

The rule selects $(\rho, \zeta) = (0, 0)$ and $\lambda_\Gamma = 1$. Table 14 reports the independent comparison, and Figure 4 shows its learning trajectory. Independent confirmation reduces cumulative regret from 1836.7 to 1535.9, a difference of 300.8 [272.2, 327.4], or 16.4% of the omission mean. Final value increases by 0.105 [0.095, 0.115]. Thus the prespecified confirmation criterion is met. Crucially, this condition is effectively fully observed: the current signal reveals the freshly sampled state, $V^* = 1$, and no past information is needed. It demonstrates a benefit from suppressing unnecessary history under this coefficient and budget; it does not repair the original recall failures, establish superiority on the original E3 setting, or validate long-memory learning.

Table 14: Independent confirmation: means and 95% seed-bootstrap intervals from 20 paired seeds. Gradient norms and finite interventions are secondary local diagnostics of the deployed terminal policy.

Metric	Omission ($\lambda_\Gamma = 0$)	Inclusion ($\lambda_\Gamma = 1$)
Cumulative regret	1836.7 [1811.9, 1862.1]	1535.9 [1507.5, 1564.4]
Final value	0.750 [0.742, 0.757]	0.855 [0.847, 0.863]
Past distractor gradient	0.073 [0.064, 0.082]	0.005 [0.004, 0.005]
Past signal gradient	0.065 [0.055, 0.074]	0.003 [0.003, 0.003]
Current distractor gradient	0.223 [0.217, 0.230]	0.161 [0.151, 0.172]
Current signal gradient	0.292 [0.285, 0.300]	0.312 [0.299, 0.326]
Past distractor flip TV	0.112 [0.096, 0.128]	0.006 [0.006, 0.007]

A diagnostic fixed before confirmation uses 256 independent episodes per trained actor. We differentiate its terminal action-one probability and report Euclidean gradient norms over past/current signal and distractor coordinates. We also flip all preterminal distractor signs, holding signals, actions,

rewards, and the current observation fixed. This stays within the binary observation support and preserves the optimal belief. The mean total-variation change in the binary action distribution falls from 0.112 to 0.006. These measurements are consistent with reduced irrelevant-history dependence while current-signal sensitivity is retained. They are neither a global certificate nor proof of causal mediation. Because sign flipping preserves the distractor distribution, the absolute policy change is not a population reward-loss estimate.

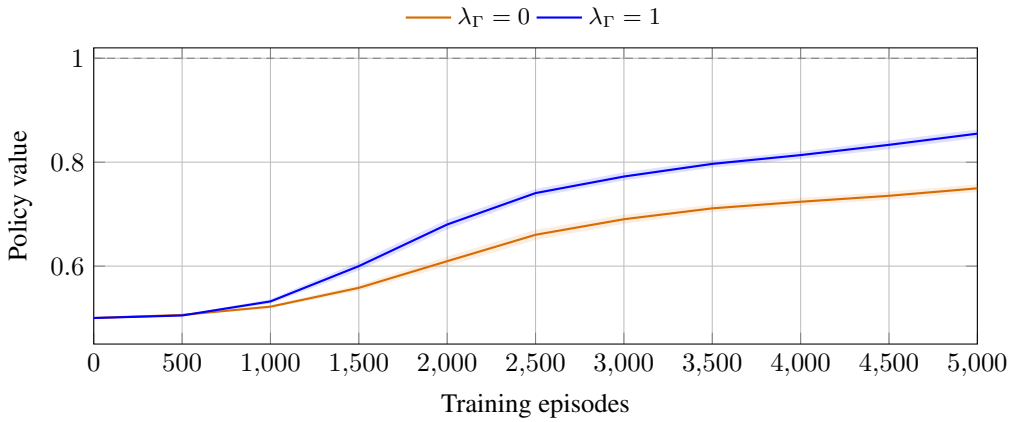


Figure 4: Independent fast-refresh confirmation, 20 paired seeds: mean value and pointwise 95% bootstrap bands, shown every tenth update. The final point evaluates the final policy; gray denotes optimal value.

All 45 pilot and 40 confirmation runs completed on HPCC A100 accelerators. The retained `results6_regime` records contain the fixed protocol, selection, individual runs, aggregates, and confirmation decision. These 85 runs are additional to the 85 initial supplementary diagnostics.

B Relation to Prior Work and Scope of the Claims

Table 15 separates established analytical tools from the control questions studied here. The contribution combines these tools under explicit conditions rather than replacing their underlying theory.

Table 15: Closest lines of work and the scope of this analysis. The underlying Jacobian, contraction, and mirror-descent tools are established.

Line of work	Established object	Present focus or distinction
Can et al. (2020); Livi (2026)	Gated recurrent dynamics and Jacobian analysis	Finite history perturbations and conditional control implications
Miller & Hardt (2019)	Finite-context approximation of stable recurrent models	Input-entry tail, decoder calibration, and control propagation
Lafuente-Mercado et al. (2025)	Local magnitude-weighted temporal influence	Uniform finite-perturbation envelope; sampled derivatives remain local
Subramanian et al. (2022); Kara & Yüksel (2022)	Information-state approximation and finite-memory control	Sensitivity as one component of a conditional representation bound
Abdolmaleki et al. (2018); Sherman et al. (2025)	Teacher-based improvement and policy mirror descent	Deployed-policy log-ratio transfer with a nonconvex recurrent actor
Cayci & Eryilmaz (2025)	Finite-time and finite-width recurrent policy-gradient analysis	Coupled LSTM envelopes; no unconditional Adam guarantee

Table 16 links each statement to its additional assumptions and applicable tasks. Filter stability alone does not verify decoder calibration or a global network envelope; the proofs below retain these requirements explicitly.

Table 16: Scope of the mathematical statements. An environmental condition does not verify learned calibration or global network envelopes.

Statement	Additional requirement	Relevant tasks
Proposition 1	Bounds on all comparison segments	E0–E4
Theorem 5	Invariant belief domain and filter contraction	E3
Corollary 6	Admissible representatives, calibration, decoder Lipschitz bound	E3, if verified
Corollary 9	Forced preterminal actions, Lipschitz belief head	E1–E3
Theorem 10	Admissible support, critic and projection errors	E1–E4
Proposition 16	Fixed certified class, independent episodes, bounded returns, approximation and optimization control	Conditional only

C Proofs

C.1 Full-path memory bounds

To prove Proposition 1, write $\Delta Z_h = Z_h - Z'_h$ and $\Delta X_h = X_h - X'_h$. Add and subtract $\mathcal{F}_\theta(Z'_{h-1}, X_h)$. Applying the fundamental theorem of calculus on the two admissible line segments gives averaged Jacobians $\widetilde{J}_h^{ZZ}, \widetilde{J}_h^{ZX}$ such that

$$\Delta Z_h = \widetilde{J}_h^{ZZ} \Delta Z_{h-1} + \widetilde{J}_h^{ZX} \Delta X_h, \qquad \|\widetilde{J}_h^{ZZ}\|_{\mathrm{op}} \le \alpha_h, \quad \|\widetilde{J}_h^{ZX}\|_{\mathrm{op}} \le \beta_h.$$

Consequently,

$$\|\Delta Z_h\|_2 \le \alpha_h \|\Delta Z_{h-1}\|_2 + \beta_h \|\Delta X_h\|_2.$$

Iterating from h to 1 yields

$$\|\Delta Z_h\|_2 \le \left(\prod_{j=1}^{h} \alpha_j \right) \|\Delta Z_0\|_2 + \sum_{s=1}^{h} \beta_s \left(\prod_{j=s+1}^{h} \alpha_j \right) \|\Delta X_s\|_2,$$

This proves the claim.

For Corollary 3, the first inequality follows directly from $\Gamma_{h,s} = \beta_s \prod_{j=s+1}^{h} \alpha_j$. For inputs outside a recent window of length m,

$$T_h(m) \le \bar{\beta} \sum_{\ell=m}^{\infty} \bar{\alpha}^{\ell} = \frac{\bar{\beta}\bar{\alpha}^m}{1-\bar{\alpha}}.$$

Solving $D_X \bar{\beta}\bar{\alpha}^m/(1-\bar{\alpha}) \le \epsilon$ for a nonnegative integer m gives

$$m_{\text{eff}}(h;\epsilon) \le \max\left\{0, \left\lceil \frac{\log\{D_X\bar{\beta}/[\epsilon(1-\bar{\alpha})]\}}{|\log\bar{\alpha}|} \right\rceil \right\}.$$

If $D_X\bar{\beta} = 0$, then $m_{\text{eff}}(h;\epsilon) = 0$; an identically zero recurrent factor can be upper-bounded by any $\bar{\alpha} \in (0,1)$ in the display above.

Contraction is sufficient, not necessary; products of norms can also be loose under rotation or cancellation.

From state influence to output influence. For a differentiable actor readout with $\|\partial \ell_h^A/\partial Z_h\|_{\text{op}} \le L_{\ell,h}$, the chain rule gives the one-way implication

$$\left\| \frac{\partial \ell_h^A}{\partial X_s} \right\|_{\text{op}} \le L_{\ell,h} \left\| J_h^{ZZ} \cdots J_{s+1}^{ZZ} J_s^{ZX} \right\|_{\text{op}} \le L_{\ell,h}\Gamma_{h,s}.$$

The converse need not hold: a small sampled input-to-logit derivative can result from readout insensitivity or cancellation while the recurrent state remains sensitive. Therefore the implemented logit-influence penalty is a local diagnostic surrogate, not an upper bound on the memory tail.

C.2 Network sensitivity, parameter envelopes, and statistical scope

Let θ contain all LSTM matrices and biases. For class-uniform bounds, take every Jacobian supremum over $\vartheta \in \text{co}(\Theta)$ and the declared domains. With $M_h = \partial \mathcal{F}_\theta/\partial\theta$, $\nu_h = \sup \|M_h\|_{\text{op}}$, and fixed Z_0, $P_h^\theta := \partial Z_h/\partial\theta = M_h + J_h^{ZZ} P_{h-1}^\theta$, where $P_0^\theta = 0$. Thus

$$\|P_h^\theta\|_{\text{op}} \le S_{\theta,h} := \sum_{s=1}^{h} \nu_s \prod_{j=s+1}^{h} \alpha_j. \tag{16}$$

This recurrent term alone does not control a trainable decoder or readout. For a scalar output coordinate, including a critic action coordinate, write

$$f_{\vartheta,h}(H_h, a) = q_{w,h}\{D_\psi(U_h^\theta), a\}, \qquad \vartheta = (\theta, \psi, w) \in \Theta \subset \mathbb{R}^p.$$

Define the global full-network sensitivity

$$K_{\text{net}}^Q := \sup_{\substack{h \le H,\ H_h \in \mathcal{H}_h,\ a \in \mathcal{A} \\ \vartheta \in \text{co}(\Theta)}} \|\nabla_\vartheta f_{\vartheta,h}(H_h, a)\|_2. \tag{17}$$

Assume $\Theta \subset \mathbb{R}^p$ has Euclidean diameter R. For finite-dimensional decoder outputs, let $L_{D,2}, Q_D$ bound its input and parameter Jacobians in $2 \to 2$ norm, and L_q, Q_w the head's Euclidean derivatives, uniformly on the declared domains. For certified $\alpha_h^+ \ge \alpha_h, \nu_h^+ \ge \nu_h$, define $S_{\theta,0}^+ = 0$ and $S_{\theta,h}^+ = \nu_h^+ + \alpha_h^+ S_{\theta,h-1}^+$.

The chain rule gives the analytic certificate

$$K_{\text{net}}^Q \le K_{\text{net}}^{Q,+} := \max_{h \le H} \left\{ (L_q L_{D,2} S_{\theta,h}^+)^2 + (L_q Q_D)^2 + Q_w^2 \right\}^{1/2}. \tag{18}$$

For vector actor logits, define analogously

$$K_{\text{net}}^A := \sup_{\substack{h \le H,\ H_h \in \mathcal{H}_h \\ \vartheta^A \in \text{co}(\Theta^A)}} \|J_{\vartheta^A} \ell_{\vartheta^A,h}(H_h)\|_{\text{op}} \le K_{\text{net}}^{A,+}.$$

The actor formula, blockwise choices of ν_h^+ and α_h^+, and the bias and hidden-state factors are given in Section C.2 and Proposition 12.

Autodifferentiated norms at fitted parameters and sampled histories are local diagnostics, not global certificates. Proposition 11 bounds sequential complexity by order $RK_{\mathrm{net}}^{Q,+}\sqrt{p\log(n)/n}$. A critic error rate additionally requires target, coverage, approximation, and optimization control; this capacity bound cannot be inverted to estimate sensitivity.

To establish the full-network complexity bound, let $\mathcal{F}_{\mathrm{read}} = \{(h, H_h, a) \mapsto f_{\vartheta,h}(H_h, a) : \vartheta \in \Theta\}$.

Proposition 11 (Full-network sequential complexity)**.** *Suppose $R > 0$, the scalar outputs are bounded, and the derivative bounds in Section 3.2 hold on parameter segments in $\mathrm{co}(\Theta)$, uniformly over nodes of admissible predictable input trees. Then*

$$\mathfrak{R}_n^{\mathrm{seq}}(\mathcal{F}_{\mathrm{read}}) \le \frac{RK_{\mathrm{net}}^{Q,+}}{n} + RK_{\mathrm{net}}^{Q,+}\sqrt{\frac{2p\log(1+2n)}{n}}.$$

If $R = 0$, the class is a singleton and the centered complexity is zero.

Differentiating the recurrence with respect to the shared recurrent parameters gives

$$\frac{\partial Z_h}{\partial \theta} = M_h + J_h^{ZZ}\frac{\partial Z_{h-1}}{\partial \theta}.$$

Unrolling and taking norms yields

$$\left\|\frac{\partial Z_h}{\partial \theta}\right\|_{\mathrm{op}} \le \sum_{s=1}^{h}\nu_s\prod_{j=s+1}^{h}\alpha_j = S_{\theta,h}.$$

The same induction with the certified bounds α_h^+, ν_h^+ gives $\|\partial Z_h/\partial\theta\|_{\mathrm{op}} \le S_{\theta,h}^+$. The coordinate projection from Z_h to U_h has operator norm one. Since θ, ψ, w are disjoint Euclidean parameter blocks, the decoder and head chain rules give

$$\|\nabla_\vartheta f_{\vartheta,h}(H_h, a)\|_2^2 \le (L_q L_{D,2} S_{\theta,h}^+)^2 + (L_q Q_D)^2 + Q_w^2 \le (K_{\mathrm{net}}^{Q,+})^2.$$

Taking the supremum proves $K_{\mathrm{net}}^{Q} \le K_{\mathrm{net}}^{Q,+}$. The mean-value theorem on parameter segments therefore implies

$$|f_{\vartheta,h}(H_h, a) - f_{\vartheta',h}(H_h, a)| \le K_{\mathrm{net}}^{Q}\|\vartheta - \vartheta'\|_2$$

at every node of every admissible predictable tree.

For completeness, the sequential complexity is the supremum over predictable input trees $\mathbf{z}$ of

$$\mathfrak{R}_n^{\mathrm{seq}}(\mathcal{F}_{\mathrm{read}}) = \sup_{\mathbf{z}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_{\mathrm{read}}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f\{\mathbf{z}_t(\epsilon_{1:t-1})\}\right].$$

Let $\mathcal{N}_r$ be an r-net of Θ. Because Θ has diameter at most R, it is contained in a radius-R ball around any of its members, and the net may be chosen with $|\mathcal{N}_r| \le (1 + 2R/r)^p$. Conditional on the past, the contribution of a fixed centering function has zero expectation against the next Rademacher sign. The sequential finite-class maximal inequality applied to the centered net, followed by the $K_{\mathrm{net}}^{Q} r$ approximation bound, gives

$$\mathfrak{R}_n^{\mathrm{seq}}(\mathcal{F}_{\mathrm{read}}) \le K_{\mathrm{net}}^{Q} r + RK_{\mathrm{net}}^{Q}\sqrt{\frac{2p\log(1+2R/r)}{n}}.$$

Taking $r = R/n$ and using $K_{\mathrm{net}}^{Q} \le K_{\mathrm{net}}^{Q,+}$ proves the displayed statement.

For the corresponding actor-logit certificate, let the actor recurrent-core bounds be $\alpha_h^{A,+}, \nu_h^{A,+}$, its decoder bounds be $L_{D,2}^A, Q_D^A$, and its logit-head bounds be L_ℓ, Q_ℓ. Then

$$S_{\theta^A,0}^+ = 0, \qquad S_{\theta^A,h}^+ = \nu_h^{A,+} + \alpha_h^{A,+}S_{\theta^A,h-1}^+,$$

$$K_{\mathrm{net}}^{A,+} := \max_{h\le H}\left\{(L_\ell L_{D,2}^A S_{\theta^A,h}^+)^2 + (L_\ell Q_D^A)^2 + Q_\ell^2\right\}^{1/2}.$$

The scalar-critic chain-rule argument also gives $K_{\mathrm{net}}^{A} \le K_{\mathrm{net}}^{A,+}$.

Proposition 12 (Explicit LSTM parameter envelope)**.** *Use the concatenated gate blocks* $\Theta_g = [W_g\ U_g\ b_g]$, $g \in \{F, I, O, G\}$, *with parameter norm* $\|\theta\|_2^2 = \sum_g \|\Theta_g\|_F^2$. *Suppose, uniformly over every admissible history and every recurrent trajectory induced by* $\vartheta \in \mathrm{co}(\Theta)$,

$$\|X_h\|_2 \le B_X, \quad \|U_{h-1}\|_2 \le B_{U,h-1}, \quad \|C_{h-1}\|_\infty \le B_{C,h-1},$$

and let $\rho_\sigma = \sup_z |\sigma'(z)| \le 1/4$. *Then a valid uniform one-step certificate is*

$$\nu_h^+ = B_{\mathrm{aug},h}\left\{2 + \rho_\sigma^2(2B_{C,h-1}^2 + 3)\right\}^{1/2}, \qquad B_{\mathrm{aug},h}^2 = B_X^2 + B_{U,h-1}^2 + 1. \tag{19}$$

The $+1$ *in* $B_{\mathrm{aug},h}^2$ *is the bias feature appearing in each gate block. Furthermore, let* $u_g^+ = \sup_{\vartheta\in\mathrm{co}(\Theta)} \|U_g\|_{\mathrm{op}}$, $f_h^+ = \sup\|F_h\|_\infty$, $o_h^+ = \sup\|O_h\|_\infty$, *and define*

$$L_{C,h}^+ = B_{C,h-1}\rho_\sigma u_F^+ + \rho_\sigma u_I^+ + u_G^+, \qquad L_{U,h}^+ = \rho_\sigma u_O^+ + o_h^+ L_{C,h}^+.$$

Then a compatible recurrent-state certificate is

$$\alpha_h^+ = \left\| \begin{bmatrix} f_h^+ & L_{C,h}^+ \\ o_h^+ f_h^+ & L_{U,h}^+ \end{bmatrix} \right\|_{\mathrm{op}}. \tag{20}$$

At fixed Z_{h-1}, X_h, put

$$\widetilde{X}_h = (X_h^\top, U_{h-1}^\top, 1)^\top, \qquad \Delta\Theta_g = [\Delta W_g\ \Delta U_g\ \Delta b_g].$$

The final coordinate accounts for biases. Direct parameter differentials satisfy

$$\begin{aligned}
\Delta_\theta C_h &= D(C_{h-1})D\big(F_h \odot (1-F_h)\big)\Delta\Theta_F\widetilde{X}_h \\
&\quad + D(G_h)D\big(I_h \odot (1-I_h)\big)\Delta\Theta_I\widetilde{X}_h \\
&\quad + D(I_h)D(1-G_h^2)\Delta\Theta_G\widetilde{X}_h, \\
\Delta_\theta U_h &= D\big(\tanh(C_h)\big)D\big(O_h \odot (1-O_h)\big)\Delta\Theta_O\widetilde{X}_h \\
&\quad + D(O_h)D\big(1-\tanh^2(C_h)\big)\Delta_\theta C_h.
\end{aligned}$$

For every block,

$$\|\Delta\Theta_g\widetilde{X}_h\|_2 \le B_{\mathrm{aug},h}\|\Delta\Theta_g\|_F.$$

Using gate and hyperbolic-tangent bounds by one, together with $\|\sigma'\|_\infty \le \rho_\sigma$, the four block coefficients for the joint perturbation $(\Delta_\theta C_h, \Delta_\theta U_h)$ are at most

$$\sqrt{2}B_{C,h-1}\rho_\sigma B_{\mathrm{aug},h}, \quad \sqrt{2}\rho_\sigma B_{\mathrm{aug},h}, \quad \rho_\sigma B_{\mathrm{aug},h}, \quad \sqrt{2}B_{\mathrm{aug},h}$$

for F, I, O, G, respectively. Cauchy–Schwarz over the four disjoint parameter blocks therefore gives

$$\|M_h\|_{\mathrm{op}}^2 \le B_{\mathrm{aug},h}^2\{2\rho_\sigma^2 B_{C,h-1}^2 + 3\rho_\sigma^2 + 2\},$$

which is Equation (19). The factor $B_{U,h-1}$ inside $B_{\mathrm{aug},h}$ is necessary because

$$\|(\Delta U_g)U_{h-1}\|_2 \le \|\Delta U_g\|_F\|U_{h-1}\|_2.$$

An entrywise hidden-state bound therefore does not eliminate this Euclidean factor.

Writing J_h^{ZZ} in cell/hidden blocks, Equation (2) gives

$$\|J_h^{CC}\|_{\mathrm{op}} \le f_h^+, \quad \|J_h^{CU}\|_{\mathrm{op}} \le L_{C,h}^+, \quad \|J_h^{UC}\|_{\mathrm{op}} \le o_h^+ f_h^+, \quad \|J_h^{UU}\|_{\mathrm{op}} \le L_{U,h}^+.$$

The standard block-norm comparison therefore gives $\|J_h^{ZZ}\|_{\mathrm{op}} \le \alpha_h^+$, proving Equation (20).

C.3 Information states and learned components

Fix a stage h and let $W_h = (Y_{h+1:H}, R_{h:H})$. Write $\mathsf{P}_h^{a_{h:H}}(\cdot \mid H_h)$ for its conditional law when the future actions are externally fixed to an admissible sequence $a_{h:H}$. This denotes an intervention on actions, not conditioning on future actions chosen by a history-dependent behavior policy. A

representation $z_h = z(H_h)$ is predictively sufficient if there exist measurable kernels G_h such that, for all admissible histories and action sequences,

$$\mathsf{P}_h^{a_{h:H}}(\cdot \mid H_h) = \mathsf{G}_h(\cdot \mid z(H_h), a_{h:H}).$$

Thus histories with the same representation have the same future observable laws. This condition can hold without identifying the latent-state belief: if two latent states have identical future observation and reward laws for every action sequence, predictions cannot distinguish beliefs over those states. Belief recovery is therefore an additional assumption, not a consequence of predictive sufficiency.

For a belief-level interpretation of either predictive variant, assume a measurable map $B_h : \mathbb{R}^{d_D} \to \mathcal{B}$ with a controlled reconstruction error

$$\sup_{H_h \in \mathcal{H}_h} \mathrm{TV}\{B_h(z(H_h)), b_h(H_h)\} \le \epsilon_{B,h}.$$

The exact belief-sufficient case has $\epsilon_{B,h} = 0$. Define the composite decoder $\widetilde{D}_h = B_h \circ D_\psi$; its belief-domain invariance, Lipschitz constants, calibration error on representative histories, and initial and one-step filtering defects must satisfy the hypotheses used in Theorem 5 and Corollary 6, with the stage index retained when evaluating successive composite decoders. In particular, the memory-to-belief bound uses calibration of $\widetilde{D}_h$, not merely predictive likelihood. For the control bound, also require

$$g_{w,h}(z) = \bar{\pi}_{w,h}\{B_h(z)\}$$

on the reachable predictive states, where $g_{w,h}$ is the GAMPO-P action head and $\bar{\pi}_{w,h}$ satisfies Assumption 7. Then Theorem 8 applies with $\hat{b}_h = \widetilde{D}_h(U_h)$, using either $\epsilon_{B,h}$ or a verified filtering bound on the reconstruction error. These conditions make explicit the additional structure needed to compare the predictive actor with an exact-belief actor using the same policy head.

Neither Equation (26) nor its terminal-inclusive extension in Equation (27) establishes equality of all interventional future laws, uniform belief reconstruction, or the policy factorization. Conditioning on adaptively chosen future actions can reveal information about intervening observations; it is not equivalent to fixing those actions by intervention. Consequently, the reported GAMPO-P results are empirical comparisons, not certifications of these assumptions. Its use as a model-free algorithm and the history-space regret decomposition remain distinct from a belief-level interpretation.

The derivatives below extend parameter sensitivity to learned inputs and initial states; they do not guarantee that an encoder preserves filtering information. For encoded-history filtering statements, a sufficient condition is a measurable left inverse of the encoder on admissible raw tuples. One-hot categorical encoding and injective affine numerical scaling satisfy this condition. A compressive learned encoder may not: if it identifies two tuples producing different Bayes updates, the encoded suffix does not specify the filter recursion. Such an encoder needs a separate sufficiency argument or an explicit representation-error term. The raw-suffix version of Corollary 6 still compares histories with identical raw suffixes; it must not be extended to arbitrary equal encoded suffixes without this check.

If $X_s = e_\eta(u_s)$ and $Z_0 = z_\zeta^0$ are learned, choose certified bounds $\beta_s^+, \kappa_s^+, \iota^+$ and add

$$S_{\eta,h}^+ = \sum_{s=1}^{h} \beta_s^+ \kappa_s^+ \prod_{j=s+1}^{h} \alpha_j^+, \qquad S_{\zeta,h}^+ = \iota^+ \prod_{j=1}^{h} \alpha_j^+,$$

where $\|J_s^{ZX}\|_{\mathrm{op}} \le \beta_s^+$, $\|\partial e_\eta/\partial \eta\|_{\mathrm{op}} \le \kappa_s^+$, and $\|\partial z_\zeta^0/\partial \zeta\|_{\mathrm{op}} \le \iota^+$ uniformly on the certified domains. For disjoint parameter blocks, the first term of Equation (18) then becomes

$$(L_q L_{D,2})^2 \{(S_{\theta,h}^+)^2 + (S_{\eta,h}^+)^2 + (S_{\zeta,h}^+)^2\}.$$

If parameters are shared, the corresponding Jacobian contributions must instead be added before taking the norm.

C.4 Belief and control error propagation

To prove Theorem 5, suppress the trajectory index and abbreviate the realized Bayes operator by Φ_h. The triangle inequality, Equation (3), and Assumption 4 imply

$$\mathrm{TV}(\hat{b}_h, b_h) \le \mathrm{TV}\{\hat{b}_h, \Phi_h(\hat{b}_{h-1})\} + \mathrm{TV}\{\Phi_h(\hat{b}_{h-1}), \Phi_h(b_{h-1})\} \le \delta_h + \rho_B \,\mathrm{TV}(\hat{b}_{h-1}, b_{h-1}).$$

Iterating this scalar recursion proves the theorem.

Uniform one-step contraction is a substantive restriction. With an unchanged latent state and uninformative observations and rewards, the Bayes update can be the identity, so its TV coefficient is one. If the filter is only nonexpansive on the same invariant domain, the same recursion instead gives

$$\mathrm{TV}(\hat{b}_h, b_h) \le \min\left\{1, e_1 + \sum_{s=2}^{h} \delta_s\right\}.$$

The horizon-independent defect bound and geometric forgetting of the initial belief then need not hold. If updates can expand TV, their Lipschitz factors must be retained in the recursion; the nonexpansive bound is not automatic. Our E3 construction satisfies strict contraction, whereas delayed recall and active sensing are not covered by this assumption. The full-path perturbation bound and the history-space regret decomposition do not require filter contraction; the control-error bound remains conditional on a valid belief error bound.

For Corollary 6, let $\widetilde{H}_h^{(m)}$ be the representative history, with recurrent state $\widetilde{Z}_h^{(m)}$ and exact belief $\widetilde{b}_h^{(m)}$. By the triangle inequality,

$$\begin{aligned}\mathrm{TV}\{D_\psi(U_h), b_h\} \le{}& \mathrm{TV}\{D_\psi(U_h), D_\psi(\widetilde{U}_h^{(m)})\} \\ &+ \mathrm{TV}\{D_\psi(\widetilde{U}_h^{(m)}), \widetilde{b}_h^{(m)}\} + \mathrm{TV}(\widetilde{b}_h^{(m)}, b_h).\end{aligned}$$

The three terms are bounded by $L_{D,\mathrm{TV}} D_X T_h(m)$, ϵ_{dec}, and $C_B \rho_B^m$, respectively.

The action bound in Theorem 8 is immediate from Assumption 7. Couple the two controlled processes so that, while their histories agree, their actions at stage s disagree with probability at most $L_\pi e_s$. Once a disagreement occurs, upper-bound all later disagreement probabilities by one. A union bound gives

$$\mathrm{TV}(\bar{d}_h^{\hat{\pi}}, \bar{d}_h^{\pi^b}) \le L_\pi \sum_{s=1}^{h-1} e_s.$$

The state–history–action distributions at stage h therefore differ by at most $L_\pi \sum_{s=1}^{h} e_s$. Since the stage reward lies in $[0,1]$,

$$|V^{\hat{\pi}} - V^{\pi^b}| \le L_\pi \sum_{h=1}^{H} \sum_{s=1}^{h} e_s = L_\pi \sum_{s=1}^{H} (H - s + 1) e_s.$$

To prove Corollary 9, forced preterminal actions give both policies the same terminal history law. At a fixed history, the expected terminal reward as a function of action lies in $[0,1]$. Its expectations under the two action laws differ by at most their TV distance. Apply Assumption 7, cap by one, and integrate. The final inequality follows from concavity of $x \mapsto \min\{1, x\}$. For E1–E3, only the terminal estimated-advantage oscillation in Theorem 10 is nonzero; clipping there to $[0,1]$ gives $(V_H^{\mathrm{osc}})^2 \le 1$.

C.5 Policy projection and regret

For any action distributions q, r, p satisfying $q \ll r$ and $q \ll p$, with $r(a)/p(a)$ finite on the support of q,

$$\mathrm{KL}(q\|p) - \mathrm{KL}(q\|r) = \sum_a q(a) \log \frac{r(a)}{p(a)} \le \max_a \left[\log \frac{r(a)}{p(a)}\right]_+.$$

Apply this identity at every history with $q = \pi_h^\star$, $r = \widetilde{\pi}_{k+1,h}$, and $p = \pi_{k+1,h}$, then take expectations and sum over stages. This proves Equation (13).

To prove Theorem 10, we view the POMDP as an MDP whose observable state is the complete history. The performance-difference identity gives

$$m_k(V^{\pi^\star} - V^{\pi_k}) = m_k \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^\star}} \langle A_h^{\pi_k}, \pi_h^\star - \pi_{k,h} \rangle.$$

After replacing $A_h^{\pi_k}$ by $\hat{A}_{k,h}$, the absolute remainder is at most $\epsilon_{Q,k}$.

Fix a history and write $q = \pi_h^\star$, $p = \pi_{k,h}$, and $\hat{A} = \hat{A}_{k,h}$. The exponential teacher satisfies the exact identity

$$\eta m_k \langle q, \hat{A} \rangle = \mathrm{KL}(q\|p) - \mathrm{KL}(q\|\widetilde{\pi}_{k+1,h}) + \log \mathbb{E}_{a\sim p} \exp\{\eta m_k \hat{A}(a)\}.$$

Because $\hat{A}$ is centered under p, $\langle p, \hat{A} \rangle = 0$. Hoeffding's lemma and the oscillation bound give

$$m_k \langle \hat{A}, q - p \rangle \le \frac{\mathrm{KL}(q\|p) - \mathrm{KL}(q\|\widetilde{\pi}_{k+1,h})}{\eta} + \frac{\eta m_k^2 (V_h^{\mathrm{osc}})^2}{8}.$$

Integrate under $d_h^{\pi^\star}$, sum over h, and use Equation (13) to replace the teacher potential by the next-policy potential at cost ξ_k/η. Summing over k telescopes Ψ_k; the final potential is nonnegative. Adding the critic remainders proves Equation (15).

Lemma 13 (Population projection bridge)**.** *Suppose $\pi_{k+1,h}(a \mid H_h) \ge \underline{p}_k > 0$, reference distributions satisfy $dd_h^{\pi^\star}/d\mu_{k,h} \le C_{P,k}$, and*

$$\epsilon_{P,k}^{\mathrm{pop}} = \sum_h \mathbb{E}_{\mu_{k,h}} \mathrm{KL}\{\pi_{k+1,h}(\cdot \mid H_h)\|\widetilde{\pi}_{k+1,h}(\cdot \mid H_h)\}.$$

Define

$$\Delta_{P,k} = \sum_h \mathbb{E}_{d_h^{\pi^\star}} \left[\mathrm{KL}(\pi_h^\star\|\pi_{k+1,h}) - \mathrm{KL}(\pi_h^\star\|\widetilde{\pi}_{k+1,h})\right].$$

Then

$$\Delta_{P,k} \le \frac{\sqrt{2C_{P,k} H \epsilon_{P,k}^{\mathrm{pop}}}}{\underline{p}_k}.$$

This is a comparator-dependent alternative to Equation (11); it does not upper-bound the comparator-free supremum ξ_k.

To prove Lemma 13, fix a history and put $p = \pi_{k+1,h}$ and $r = \widetilde{\pi}_{k+1,h}$. If $p(a) \ge \underline{p}_k$, then

$$\left[\sum_a \pi_h^\star(a) \log \frac{r(a)}{p(a)}\right]_+ \le \frac{\|r - p\|_1}{\underline{p}_k}.$$

Let $g_h(H_h) = \|r - p\|_1$ and $w_h = dd_h^{\pi^\star}/d\mu_{k,h}$. Because $w_h \le C_{P,k}$ and $\mathbb{E}_{\mu_{k,h}} w_h = 1$,

$$\mathbb{E}_{d_h^{\pi^\star}} g_h \le \sqrt{C_{P,k} \mathbb{E}_{\mu_{k,h}} g_h^2} \le \sqrt{2C_{P,k} \mathbb{E}_{\mu_{k,h}} \mathrm{KL}(p\|r)},$$

where the last step is Pinsker's inequality. Summing over stages and using Cauchy–Schwarz yields

$$\sum_h \mathbb{E}_{d_h^{\pi^\star}} g_h \le \sqrt{2C_{P,k} H \epsilon_{P,k}^{\mathrm{pop}}},$$

which proves the claim.

Lemma 14 (Exploration and potential transfer)**.** *Let $\pi_{k+1,h} = (1 - \gamma_{k+1})\bar{\pi}_{k+1,h} + \gamma_{k+1} u_h$, with $0 \le \gamma_{k+1} < 1$. Let $\bar{\xi}_k$ be Equation (11) with the fitted actor in the denominator. If at most H_{dec} stages admit more than one action,*

$$\xi_k \le \bar{\xi}_k + H_{\mathrm{dec}} \log \frac{1}{1 - \gamma_{k+1}}.$$

The same inequality holds for maxima over any fixed sample.

Proof. The deployed action probability is at least $(1 - \gamma_{k+1})\bar{\pi}_{k+1,h}(a \mid H_h)$. Take log ratios and use $[x + c]_+ \le [x]_+ + c$ for $c \ge 0$. Forced-action stages have zero distortion. Sum over the remaining stages. □

This potential-transfer correction differs from the value cost of exploration. For $\gamma_k \asymp k^{-1/2}$, its cumulative upper bound is $O(\sqrt{K})$; dividing it by $\eta \asymp K^{-1/2}$ can yield a linear regret term. The exploration value bound alone therefore does not close the deployed-policy telescope. Direct control of deployed distortion can be substantially sharper.

Lemma 15 (Critic risk under history coverage)**.** *At epoch k, suppose $dd_h^{\pi^\star}/dd_h^{\pi_k} \le C_{d,k}$ and $\pi_{k,h}(a \mid H_h) \ge p_k > 0$ on admissible actions. Define*

$$r_{Q,k} = \frac{1}{H}\sum_h \mathbb{E}_{H_h\sim d_h^{\pi_k},\, a\sim\pi_{k,h}}\{\widehat{Q}_{k,h}(H_h,a) - Q_h^{\pi_k}(H_h,a)\}^2.$$

Then

$$\epsilon_{Q,k} \le m_k H\sqrt{C_{d,k} r_{Q,k}/p_k}. \tag{21}$$

Proof. Put $e(a) = \widehat{Q}_{k,h}(H_h,a) - Q_h^{\pi_k}(H_h,a)$. Centering changes e by a constant in a, whose inner product with $\pi_h^\star - \pi_{k,h}$ is zero. At any history, Cauchy–Schwarz gives

$$|\langle e, \pi_h^\star - \pi_{k,h}\rangle| \le \left(\sum_a \pi_{k,h}(a)e(a)^2\right)^{1/2}\left(\sum_a \frac{(\pi_h^\star(a)-\pi_{k,h}(a))^2}{\pi_{k,h}(a)}\right)^{1/2}.$$

The second squared factor equals $\sum_a \pi_h^\star(a)^2/\pi_{k,h}(a) - 1 \le 1/p_k$. Changing measure to $d_h^{\pi_k}$, applying Cauchy–Schwarz, and using the density-ratio bound controls the first factor's expectation by the square root of $C_{d,k}$ times its on-policy second moment. Sum over stages and apply Cauchy–Schwarz once more. □

An action-probability floor does not bound $C_{d,k}$ by a useful horizon-independent constant. Rare comparator histories can make coverage restrictive.

Proposition 16 (Finite-sample Monte Carlo critic bridge)**.** *Condition on the data preceding epoch k. Let $n_k = |\mathcal{D}_k^Q| \ge 1$ independent on-policy episodes be used in Equation (23). Suppose returns and fitted critic outputs lie in $[0,B]$, the fixed critic parameter class has dimension p_Q, diameter $R_Q > 0$, and class-uniform output Lipschitz constant G_Q. Suppose its empirical penalized objective is optimized within $\epsilon_{\mathrm{opt},k}$ of the infimum and the nonnegative gradient penalty is at most G_Q^2. Here $\lambda_{Q,k}$ includes any fixed normalization of this penalty. Let a_k^2 be the infimum of normalized on-policy squared approximation error to Q^{π_k}, and define*

$$u_k = B^2\sqrt{\frac{p_Q\log(1+2n_k)+\log(2/\delta_k)}{2n_k}} + \frac{4BR_QG_Q}{n_k}, \qquad 0<\delta_k<1.$$

With conditional probability at least $1-\delta_k$,

$$r_{Q,k} \le a_k^2 + 2u_k + \epsilon_{\mathrm{opt},k} + \lambda_{Q,k}G_Q^2. \tag{22}$$

Proof. Treat a complete episode as one observation. Its average squared loss

$$\ell_q = H^{-1}\sum_h (q_h(H_h,A_h) - G_h)^2$$

lies in $[0,B^2]$ and is $2BG_Q$-Lipschitz in the parameter. A radius-R_Q/n_k parameter net has at most $(1+2n_k)^{p_Q}$ points. Hoeffding's inequality and a union bound control the empirical-to-population gap on this net by the square-root term in u_k. Moving to the net changes each empirical and population loss by at most $2BG_QR_Q/n_k$, giving the remaining term. This argument uses independence between episodes, not between stages.

Since $\mathbb{E}[G_h \mid H_h, A_h] = Q_h^{\pi_k}(H_h,A_h)$, population squared loss equals approximation error plus a noise term independent of q. Compare the empirical penalized fit with an arbitrarily near-optimal population approximant. Two uniform-deviation terms, the optimization gap, and the approximant's penalty bound give Equation (22). □

Choose $\sum_k \delta_k \le \delta$ for a simultaneous guarantee. This result is conditional on approximation and optimization quality; it is not a convergence theorem for Adam. Fixed batch size and fixed positive regularization do not make this upper bound vanish. The supplied $m_k = 50$, 50–50 split gives $n_k = 25$; the bound is not asserted to be numerically informative for these runs.

Corollary 17 (Regret in terms of population fitting errors)**.** *Under Theorem 10 and Lemma 15, use* $\mu_{k,h} = d_h^{\pi_k}$ *in Lemma 13. Let* $p_{k+1} > 0$ *be a floor for the next deployed policy, and set*

$$D_K = \sum_{k=1}^{K} \frac{\sqrt{2C_{d,k} H \epsilon_{P,k}^{\mathrm{pop}}}}{p_{k+1}}, \quad M_K = \sum_{k=1}^{K} m_k^2, \quad S_{\mathrm{osc}} = \sum_h (V_h^{\mathrm{osc}})^2.$$

Then

$$\mathrm{Regret}_{N_K}(\pi^\star) \le \frac{\Psi_1 + D_K}{\eta} + \frac{\eta M_K S_{\mathrm{osc}}}{8} + H \sum_k m_k \sqrt{C_{d,k} r_{Q,k}/p_k}.$$

If a bound $\mathcal{B}_K \ge D_K$ *is fixed in advance and valid for the selected stepsize, with* $\Psi_1 + \mathcal{B}_K > 0$ *and* $M_K S_{\mathrm{osc}} > 0$*, choosing*

$$\eta = \sqrt{8(\Psi_1 + \mathcal{B}_K)/(M_K S_{\mathrm{osc}})}$$

gives

$$\mathrm{Regret}_{N_K}(\pi^\star) \le \sqrt{(\Psi_1 + \mathcal{B}_K) M_K S_{\mathrm{osc}}/2} + H \sum_k m_k \sqrt{C_{d,k} r_{Q,k}/p_k}.$$

Proof. In the proof of Theorem 10, bound comparator potential transfer by Lemma 13 instead of ξ_k. Apply Lemma 15, then minimize $a/\eta + b\eta$ using the prespecified upper bound $\mathcal{B}_K$. □

Projection errors depend on the stepsize through the teacher and learned policies. Substituting a realized D_K into a stepsize formula after training is therefore not an algorithmic guarantee. The required budget must be justified for the selected stepsize, or uniformly over the candidate stepsizes. For fixed H, bounded positive epoch lengths, $\mathcal{B}_K = o(K)$, and

$$\sum_k m_k \sqrt{C_{d,k} r_{Q,k}/p_k} = o(N_K),$$

the bound is sublinear. The empirical projection loss still needs a population bound, and these conditions have not been verified for the reported runs.

Lemma 18 (Finite-cover projection certificate)**.** *Equip each declared history domain with a metric* d_h*. Let* $\mathcal{N}_{k,h}$ *be an* $r_{k,h}$*-cover by admissible histories. Suppose the action set is constant within each covering cell, and every deployed-policy log-ratio*

$$g_{k,h,a}(H_h) = \log \frac{\widetilde{\pi}_{k+1,h}(a \mid H_h)}{\pi_{k+1,h}(a \mid H_h)}$$

is $L_{k,h}$*-Lipschitz throughout that cell. Then*

$$\xi_k \le \sum_h \max_{H_h \in \mathcal{N}_{k,h},\, a} [g_{k,h,a}(H_h)]_+ + \sum_h L_{k,h} r_{k,h}.$$

Proof. Choose a covering history within $r_{k,h}$ of any history. Apply the Lipschitz bound and the 1-Lipschitz positive-part map, then take action and history suprema and sum over h. □

A random training batch need not be a cover. Exact enumeration gives radius zero on a finite history domain but can be exponential in H. For continuous domains, finite constants require certified history sensitivity and positive probabilities. Local derivatives do not establish these conditions. The lemma specifies what would turn a sampled maximum into a certificate without asserting that the reported audits meet it.

Proposition 19 (Comparator approximation)**.** *Let* Π_A *be the recurrent actor class and* $\pi_A^\star \in \arg\max_{\pi \in \Pi_A} V^\pi$*. For any comparator,*

$$\mathrm{Regret}_{N_K}(\pi^\star) = \mathrm{Regret}_{N_K}(\pi_A^\star) + N_K\{V^{\pi^\star} - V^{\pi_A^\star}\}.$$

Suppose $\pi^\star$ *is an exact-belief policy and* Π_A *contains a policy* $\widetilde{\pi}_A$ *that applies the same Lipschitz head to decoded beliefs. Then*

$$V^{\pi^\star} - V^{\pi_A^\star} \le V^{\pi^\star} - V^{\widetilde{\pi}_A} \le L_\pi \sum_{h=1}^{H} (H - h + 1)\{\epsilon_{\mathrm{dec}} + C_B \rho_B^m + L_{D,\mathrm{TV}} D_X T_h^A(m)\}.$$

The identity follows by adding and subtracting the class-optimal value. Class optimality gives $V^{\pi^\star_A} \geq V^{\widetilde{\pi}_A}$, and the final inequality is Theorem 8 and Corollary 6.

Theorem 20 (Delayed-cue necessity)**.** *For any $H \geq 3$, consider a POMDP with a uniform latent bit $B \in \{0,1\}$, revealed at $h = 1$, identical intermediate observations, zero preterminal rewards, forced no-op actions before H, terminal actions $\{0,1\}$, and $R_H = \mathbf{1}\{A_H = B\}$. Let*

$$\epsilon_\pi = \mathrm{TV}\{\pi_H(\cdot \mid B = 0), \pi_H(\cdot \mid B = 1)\}.$$

Then $V^\pi \leq (1 + \epsilon_\pi)/2$, while $V^\star = 1$, and deploying π for N episodes incurs regret at least $N(1 - \epsilon_\pi)/2$.

If $\pi_H(\cdot \mid H_H) = g_H(Z_H)$, $\mathrm{TV}\{g_H(z), g_H(z')\} \leq L_Z\|z - z'\|_2$, and paired trajectories can be coupled so that only X_1 differs, by at most D_X, then

$$\epsilon_\pi \leq \min\{1, L_Z D_X \Gamma_{H,1}\}, \qquad \mathrm{Regret}_N(\pi^\star) \geq \frac{N}{2}\left[1 - \min\{1, L_Z D_X \Gamma_{H,1}\}\right].$$

Let $p_b = \pi_H(A_H = 1 \mid B = b)$. The expected reward is

$$\frac{1}{2}\{1 - p_0 + p_1\} \leq \frac{1}{2}(1 + |p_1 - p_0|) \leq \frac{1 + \epsilon_\pi}{2}.$$

The history-dependent policy choosing $A_H = B$ has value one. Summing the per-episode gap proves the regret bound. If the paired histories differ only at X_1, Proposition 1 gives $\|Z_H^{(1)} - Z_H^{(0)}\|_2 \leq D_X \Gamma_{H,1}$. Applying the declared L_Z-Lipschitz terminal action map proves the gate-based statement.

When $m_k = 1$ and $E_D \geq \sum_k \xi_k$, optimizing the first two terms of Equation (15) with

$$\eta = \sqrt{\frac{8(\Psi_1 + E_D)}{K \sum_h (V_h^{\mathrm{osc}})^2}}$$

gives

$$\mathrm{Regret}_K(\pi^\star) \leq \sqrt{\frac{(\Psi_1 + E_D)K}{2} \sum_{h=1}^{H} (V_h^{\mathrm{osc}})^2} + \sum_{k=1}^{K} \epsilon_{Q,k}.$$

Fixed blocks of length m produce the expected $\sqrt{m}$ factor because $\sum_k m_k^2 = m N_K$.

Suppose a deployed policy is $\pi_{k,h} = (1 - \gamma_k)\bar{\pi}_{k,h} + \gamma_k u_h$, where u_h is uniform over admissible actions. Then $\mathrm{TV}(\pi_{k,h}, \bar{\pi}_{k,h}) \leq \gamma_k$. Repeating the coupling argument used for Theorem 8,

$$|V^{\pi_k} - V^{\bar{\pi}_k}| \leq \frac{H(H+1)}{2}\gamma_k.$$

The cumulative cost is at most

$$\frac{H(H+1)}{2} \sum_{k=1}^{K} m_k \gamma_k.$$

Consequently, a fixed positive γ_k generally produces a linear term, whereas $\gamma_k \asymp k^{-1/2}$ makes this value-cost term sublinear when epoch lengths are uniformly bounded. This does not control the separate potential-transfer term in Lemma 14. Action support alone does not give a useful horizon-independent history-occupancy ratio.

D Implementation and Reproducibility

D.1 Algorithm and network configuration

Initialize actor and critic in their declared parameter domains and form a full-support deployed policy. Choose KF, P, or P-R and all schedules before evaluation. For epochs $k = 1, \ldots, K$:

1. Deploy π_k for m_k episodes and split complete trajectories into $\mathcal{D}_k^Q, \mathcal{D}_k^A$.

2. Fit the Monte Carlo critic by Equation (23), evaluate all admissible actions on actor histories, and center under deployed π_k.
3. Freeze the old policy and fitted critic while constructing the teacher in Equation (8); detach it from actor optimization.
4. Approximately minimize Equation (9) using the information loss in Equation (25), Equation (26), or Equation (27). In predictive variants, jointly optimize the prediction-head parameters χ.
5. Apply the declared parameter constraints. A certified run must additionally satisfy Section D.3; otherwise retain its diagnostic-only status.
6. Form $\pi_{k+1} = (1-\gamma_{k+1})\pi_{\vartheta_{k+1}} + \gamma_{k+1}u$. Log fitted and deployed projection distortion and use the deployed policy for the next epoch.

K counts updates and $N_K = \sum_k m_k$ counts training episodes. Regret uses $\pi_1, \dots, \pi_K$. The reported protocol evaluates the final deployed policy π_{K+1}, including the exploration mixture. This convention and the recorded pre-mixture distortions must be checked against the evaluation configuration; fitted and deployed policy results must remain distinguishable.

Categorical variables are one-hot encoded and numerical inputs are scaled to $[-1, 1]$ by a fixed encoder. The filtering interpretation requires scaling to remain injective on the declared input range: clipping or quantization must not merge distinct admissible tuples. Actor and critic use separate width-64 LSTMs, decoders, and two-layer, 64-unit tanh heads. The executed runs apply information losses to both networks, with critic coefficients 10 for KF and 1 for predictive consistency. Recurrent states are recomputed from complete trajectories after every parameter update; hidden states cached under older parameters are never reused. Within each epoch, a fixed 50–50 trajectory split gives 25 episodes to each fit. The code declares validation seeds 100–109 but disables tuning; no completed matched ten-seed search is documented.

Encoder baselines use a width-64 GRU; a one-block, single-head Transformer with width 64, feedforward width 256, and 128 learned positions; or S4D with 64 channels and eight complex modes per channel. MLP controls use windows of 16 or one observation. The predictive decoder has width 16.

D.2 Training objectives and optimization

For a fresh on-policy trajectory, let $G_{i,h} = \sum_{t=h}^{H} R_{i,t}$. The ideal clipped Monte Carlo critic fit minimizes

$$\widehat{Q}_k \in \arg\min_{q \in \mathcal{Q}(G_Q)} \frac{1}{|\mathcal{D}_k^Q| H} \sum_{i \in \mathcal{D}_k^Q} \sum_{h=1}^{H} \{q_h(H_{i,h}, A_{i,h}) - G_{i,h}\}^2 + \lambda_{Q,k} \widehat{\mathcal{R}}_{Q,k}(q), \tag{23}$$

where $\mathcal{Q}(G_Q) = \{q_\vartheta : \vartheta \in \Theta^Q(G_Q)\}$ and $\widehat{\mathcal{R}}_{Q,k}$ averages squared parameter-gradient norms at the sampled state–action pairs. Outputs are clipped to $[0, H-h+1]$, and

$$\widehat{A}_{k,h}(H_h, a) = \widehat{Q}_{k,h}(H_h, a) - \sum_{a'} \pi_{k,h}(a' \mid H_h) \widehat{Q}_{k,h}(H_h, a'). \tag{24}$$

The optimization displays specify targets; numerical training only approximately minimizes them, and the analysis retains the resulting errors. Earlier replay requires explicit off-policy correction and is not used in the primary analysis.

For GAMPO-KF, the softmax belief decoder minimizes

$$\widehat{\mathcal{L}}_{B,k}^{\mathrm{KF}} = \frac{1}{|\mathcal{D}_k^A|} \sum_i \left[\mathrm{KL}(b_{i,1} \| \widehat{b}_{i,1}) + \frac{1}{H-1} \sum_{h=2}^{H} \mathrm{KL}\Big(\Phi_{h, A_{i,h-1}, R_{i,h-1}, Y_{i,h}}(\widehat{b}_{i,h-1}) \| \widehat{b}_{i,h} \Big) \right]. \tag{25}$$

For $H = 1$, omit the empty consistency term.

For the nominal GAMPO-P block objective, let $T_{i,h,\ell} = (Y_{i,h+\ell}, R_{i,h:h+\ell-1})$, $W_{i,h,\ell} = \{D_\psi(U_{i,h}^A), A_{i,h:h+\ell-1}\}$, and $\mathcal{I}_k = \{(i,h,\ell) : i \in \mathcal{D}_k^A, \ell \in \mathcal{L}, h+\ell \le H\}$. Its teacher-forced predictive loss is

$$\widehat{\mathcal{L}}_{B,k}^{\mathrm{P}} = -\frac{1}{|\mathcal{I}_k|} \sum_{(i,h,\ell) \in \mathcal{I}_k} \log p_{\chi,\ell}(T_{i,h,\ell} \mid W_{i,h,\ell}). \tag{26}$$

Every reward block in Equation (26) ends by $H-1$. The executed implementation instead allows $h+\ell \le H+1$; at equality it predicts only rewards through R_H, without a future observation. This includes terminal supervision on actor and critic, but differs from a dedicated terminal head. In particular, the implemented $\ell = 1, h = H$ term already predicts R_H from the terminal representation using a shared one-step head. Thus GAMPO-P-R changes head sharing and loss weighting, not the presence of terminal supervision. It adds

$$\widehat{\mathcal{L}}_{\mathrm{T},k} = -\frac{1}{|\mathcal{D}_k^A|} \sum_{i \in \mathcal{D}_k^A} \log p_{\chi,\mathrm{T}}(R_{i,H} \mid D_\psi(U_{i,H}^A), A_{i,H}),$$
$$\widehat{\mathcal{L}}_{B,k}^{\mathrm{P\text{-}R}} = \widehat{\mathcal{L}}_{B,k}^{\mathrm{P}} + \lambda_{\mathrm{T}} \widehat{\mathcal{L}}_{\mathrm{T},k}. \tag{27}$$

Use a Bernoulli likelihood for the binary terminal rewards in E1–E4. The head receives the pre-action state at H, the realized action, and the reward as its target. No terminal reward enters the actor input before the action. Gradients update the actor core, decoder, and prediction heads, with a separate critic. Whenever $\mathcal{I}_k$ is empty, including $H = 1$, define the block-prediction loss to be zero. New runs must record $\lambda_{\mathrm{T}} > 0$, the prediction-head architecture, and any numerical probability floor. The RL tables use the extended block implementation, not this separate terminal-state head.

Proposition 21 (Terminal reward identifies the recall cue)**.** *For E1 with $0 \le q < 1/2$, or E2 after both cues have been observed, suppose the terminal behavior action is randomized independently of the history, with positive probability for each prediction action. If a representation z_H and binary reward head reproduce the exact conditional terminal reward law for every history and action, histories with different posterior success probabilities cannot share z_H. In E1 the representation must distinguish the two noisy cue values; in E2 it must distinguish the two terminal target values.*

Proof. In E1, $\Pr(R_H = 1 \mid H_H, A_H = 1) = \Pr(B = 1 \mid H_H)$ equals q or $1-q$ according to the cue. In E2 this probability is the observed-history target $T \in \{0, 1\}$. Two such histories with the same z_H would receive the same prediction for action 1, a contradiction. Independent action randomization prevents the conditioning action from itself revealing the cue. The Bernoulli log score is strictly proper, so an exact population optimum in a realizable class reproduces these probabilities. □

This is an ideal identification condition, not a finite-sample learning guarantee. The deployed actor's exploration mixture is generally history-dependent and does not meet the independent-randomization premise. A diagnostic bank with externally randomized terminal actions can test this condition. The nominal loss in Equation (26) cannot impose it in E1: its observation targets are independent of the cue and reward targets are zero. In E2, after the long cue has passed, its future block targets similarly do not require retention of that cue. This argument does not establish that the executed extended blocks lack terminal supervision, or identify the cause of their empirical recall failure.

The primary lag weights are $\omega_{h,s} = 1/(h-1)$ for $s < h$, with an empty sum at $h = 1$. Using the local quantities defined in Section 5,

$$\widehat{\mathcal{R}}_{\Gamma,k}^A = \frac{1}{|\mathcal{D}_k^A| H} \sum_{i,h} \sum_{s<h} \omega_{h,s} (\widehat{J}_{i,h,s}^A)^2, \qquad \widehat{\mathcal{R}}_{K,k}^A = \frac{1}{|\mathcal{D}_k^A| H} \sum_{i,h} (\widehat{K}_{i,h}^{A,\mathrm{loc}})^2.$$

For pre-mixture projection diagnostics,

$$\widehat{\bar{\xi}}_k(\vartheta) = \sum_{h=1}^{H} \max_{i \in \mathcal{D}_k^A} \max_{a \in \mathcal{A}} \left[\log \frac{\widetilde{\pi}_{k+1,h}(a \mid H_{i,h})}{\pi_{\vartheta,h}(a \mid H_{i,h})} \right]_+, \qquad \widehat{\mathcal{R}}_{D,k} = \widehat{\bar{\xi}}_k(\vartheta)^2. \tag{28}$$

A deployed-policy audit instead uses

$$\widehat{\xi}_k^{\mathrm{dep}} = \sum_h \max_{i \in \mathcal{D}_k^A,\, a} \left[\log \frac{\widetilde{\pi}_{k+1,h}(a \mid H_{i,h})}{(1-\gamma_{k+1}) \bar{\pi}_{k+1,h}(a \mid H_{i,h}) + \gamma_{k+1} u_h(a)} \right]_+. \tag{29}$$

Population ξ_k uses this deployed denominator. Lemma 14 bounds its change from fitted-actor distortion on the same histories. A fitted-actor penalty can be retained, but must not be reported as deployed distortion. Optimizing a deployed-policy penalty directly is a separate, unevaluated ablation.

Squaring the nonnegative sampled distortion increases the relative penalty on large deviations; it does not change its zero set or make the maximum smooth. Let $\bar{\xi}_k$ be the population version of $\widehat{\bar{\xi}}_k$. If a uniform gap bound $\bar{\xi}_k \le \widehat{\bar{\xi}}_k + g_k$ is established, then Equation (12) and $\widehat{\mathcal{R}}_{D,k} \le r_k$ imply

$$\xi_k \le \sqrt{r_k} + g_k + H_{\mathrm{dec}} \log \frac{1}{1-\gamma_{k+1}}.$$

Without such a gap bound, the sampled penalty supplies no population control. Lemma 18 gives a sufficient covering condition for either fitted or deployed log ratios.

The executed runs use fitted-Q targets rather than the reference Monte Carlo objective in Equation (23):

$$Y^k_{i,h} = R_{i,h} + \sum_{a'} \pi_{k,h+1}(a' \mid H_{i,h+1}) q^{\mathrm{tar}}_{k,h+1}(H_{i,h+1}, a'),$$

where a frozen copy of the complete recurrent critic produces the targets. The critic is copied at each epoch start and frozen for its 20 fitting steps. Regression uses raw outputs; teacher and target evaluations clip them to $[0, H-h+1]$. Critic MSE, sensitivity, and actor KL average over admissible decision stages: the terminal stage on E1–E3 and all stages on E4. Thus E1–E3 use terminal-return targets, whereas E4 bootstraps. Earlier-policy replay is not used.

The executed optimizer is Adam (Kingma & Ba, 2015) at learning rate 10^{-3}, with gradient norm clipped at one, 20 critic and 10 actor steps per epoch, each on its complete 25-episode split. Teacher temperature $\eta m_k = 0.3$. GAMPO-P uses $(\lambda_\Gamma, \lambda_K, \lambda_B, \lambda_D, \lambda_Q) = (0.1, 0, 0.1, 0.01, 0.001)$; GAMPO-KF changes the actor information-loss weight to 10. Thus the $\lambda_K = \lambda_Q = 0$ ablation removes only the critic penalty; the primary actor coefficient is already zero. The $\lambda_B = 0$ ablation removes information losses from both networks. Each penalty is divided by its first-call median magnitude, floored at 1, then held fixed. Code comments describe preliminary single-seed choices on E3–E4, not a matched held-out search. Comparisons assess this fixed configuration, not optimally tuned methods.

The deployed actor is mixed with the uniform admissible-action policy using $\gamma_k = 0.1/\sqrt{k}$. This supplies action coverage and full support; its value cost is bounded in Section C.5. Teacher probabilities are computed with log-sum-exp. A sensitivity-adaptive teacher temperature is an optional ablation outside the fixed-step setting of Theorem 10.

For GAMPO-P, the predictive lag set is $\mathcal{L} = \{1, 2, 4, 8, 16, 32, 64\} \cap \{1, \dots, H-1\}$. Categorical targets use softmax likelihoods and continuous targets use fixed-variance Gaussian likelihoods after scaling. The nonsmooth maxima in Equation (28) are replaced during training by log-sum-exp with temperature 0.05; evaluation uses the exact maximum. Gradients flow through the actor, decoder, and prediction heads, while future target variables and teacher-forced actions are detached.

The reported influence loss uses $\omega_{h,s} = 1/(h-1)$ for all $s < h$, regardless of decision relevance. It differs from a certified tail with chosen window m. For controlled audits, partition input coordinates into task cues and independent distractors, and evaluate their Jacobian blocks and paired finite interventions separately. This partition is an evaluation device available from the synthetic generator, not privileged learner input.

Let $\mathcal{D}_{\mathrm{dec}}$ be the decision stages and $H_{\mathrm{dec}} = |\mathcal{D}_{\mathrm{dec}}|$. The all-stage reference objective divides terminal KL by H. The executed implementation already averages KL and sampled influence over decision stages, retaining every source lag. A new all-stage versus decision-stage comparison should therefore be labeled an ablation of this executed setting, not an untested correction presumed absent from the tables.

Raw-logit influence also depends on the logit gauge: $\ell_h + c(H_h)\mathbf{1}$ gives the same policy with potentially different input derivatives. An additional ablation can use centered logits $\ell_h - \mathbf{1}(\mathbf{1}^\top \ell_h)/|\mathcal{A}_h|$ or policy-probability derivatives. The reported raw-logit diagnostics remain labeled as such. Across λ_Γ, jointly report cue influence, distractor influence, terminal reward prediction, regret, and final value.

D.3 Jacobian computation and certification

The executed penalty materializes the Jacobian of actor logits with respect to the input trajectory. Each step samples up to four admissible target stages uniformly, includes every source stage $s < h$, and

uses weights $1/(h-1)$. It does not sample four source lags per target. Two power iterations estimate the operator norm; singular vectors are detached while the final norm retains mixed derivatives. Explicit joint-state products belong to the separate mechanistic audit. The scalar critic diagnostic is computed directly by reverse-mode automatic differentiation,

$$\widehat{K}^{Q,\mathrm{loc}}_{\mathrm{net},i,h} = \max_{a\in\mathcal{A}} \left\| \nabla_{\vartheta^Q} q_{\vartheta^Q,h}(H_{i,h},a)\big|_{\vartheta^Q=\widehat{\vartheta}^Q} \right\|_2 ,$$

whereas the actor logit-Jacobian norm in Section 5 is estimated by alternating Jacobian–vector and vector–Jacobian products. Diagnostics should include iteration tolerances and residuals. Finite-step power iteration can underestimate the leading singular value, so the resulting quantity remains a numerical diagnostic.

For clarity, the pair count is not the arithmetic cost. Let c_{step} bound the work of a one-step Jacobian–vector or vector–Jacobian product, including readout costs in this bound, and let a_{step} bound saved activations per recurrent step. For dense gates the core contribution to c_{step} is $O(d_u(d_u+d_x))$. Independent path traversals cost

$$O\left(n_{\mathrm{pow}} c_{\mathrm{step}} \sum_{(h,s)\in\mathcal{I}_{\mathrm{lag}}} (h-s+1)\right)$$

per trajectory, where $\mathcal{I}_{\mathrm{lag}}$ is the set of evaluated source–target pairs. This gives the exhaustive and sampled bounds in Section 5. These are bounds for this implementation strategy, not lower bounds for all ways of sharing derivative computations. Ordinary backpropagation through time costs $O(Hc_{\mathrm{step}})$. Differentiating the influence penalty adds mixed second derivatives and larger constants. Processing sampled terms serially and recomputing their graphs can use $O(Ha_{\mathrm{step}}+p+LH)$ storage, including p parameters and optimizer state; retaining all higher-order graphs can require substantially more. Batch costs multiply by the trajectory count. These arithmetic estimates do not replace measurements of accelerator time and peak memory, which are not reported in the current tables.

For recurrent parameter transport, we also record

$$\widehat{S}^{\mathrm{env}}_{\theta,i,h} = \sum_{s=1}^{h} \|\widehat{M}_{i,s}\|_{\mathrm{op}} \prod_{j=s+1}^{h} \|\widehat{J}^{ZZ}_{i,j}\|_{\mathrm{op}}.$$

Using pointwise decoder and critic-head Jacobian norms evaluated on the same history, we name the resulting local envelope

$$\widehat{K}^{Q,\mathrm{env,loc}}_{\mathrm{net},i,h} = \left\{ (\widehat{L}_{q,i,h}\widehat{L}_{D,2,i,h}\widehat{S}^{\mathrm{env}}_{\theta,i,h})^2 + (\widehat{L}_{q,i,h}\widehat{Q}_{D,i,h})^2 + \widehat{Q}^2_{w,i,h} \right\}^{1/2}.$$

This envelope is local to the sampled history. The corresponding influence diagnostics are

$$\widehat{J}^{\mathrm{prod}}_{h,s} = \|J^{ZZ}_h J^{ZZ}_{h-1} \cdots J^{ZZ}_{s+1} J^{ZX}_s\|_{\mathrm{op}},$$

$$\widehat{J}^{\mathrm{env}}_{h,s} = \|J^{ZX}_s\|_{\mathrm{op}} \prod_{j=s+1}^{h} \|J^{ZZ}_j\|_{\mathrm{op}},$$

$$\widehat{J}^{\mathrm{FG}}_{h,s} = \|J^{ZX}_s\|_{\mathrm{op}} \prod_{j=s+1}^{h} \|D(F_j)\|_{\mathrm{op}}.$$

The first retains cancellations and changes of singular direction; the second matches the local norm-product envelope; the third retains only the direct forget route.

A reported $K^{r,+}_{\mathrm{net}}$ uses predetermined bounds on $\|X_h\|_2$, $\|U_h\|_2$, and $\|C_h\|_\infty$, together with verified recurrent, parameter-domain, decoder, and head bounds, to evaluate Equations (18) to (20). The fixed encoder gives an exact B_X; the LSTM equations give $B_{U,h} \le \sqrt{d_u}$ and, for $C_0 = 0$, $B_{C,h} \le h$. Interval propagation through the gate preactivations gives f^+_h and o^+_h.

For class certification, fix a compact convex envelope $\overline{\Theta}^r$ of norm balls and bias intervals before fitting. Verify state and derivative bounds uniformly on this envelope, with

$$\mathrm{co}\{\Theta^r(G_r)\} \subseteq \overline{\Theta}^r, \qquad \sup_{\vartheta\in\overline{\Theta}^r} K^{r,+}_{\mathrm{net}}(\vartheta) \le G_r.$$

Equivalently, evaluate a common envelope from worst allowed gate, decoder, and readout norms. An accepted iterate's smaller pointwise certificate does not replace the class-level bound. If the fixed envelope fails the threshold, redesign and verify it; shrinking one proposed step does not repair a class certificate.

A feasibility wrapper can additionally enforce candidate-specific conditions. Starting from a feasible iterate ϑ, compute the Adam proposal Δ once and test

$$\vartheta^{(j)} = \mathsf{P}_r(\vartheta + \beta^j \Delta), \qquad j = 0, \dots, J_{\max},$$

where P_r projects onto the fixed convex envelope, and $\beta \in (0,1)$, $J_{\max}$ are prespecified. Accept the first candidate meeting all declared additional conditions. If all fail, retain the previous parameters and optimizer state, including the step counter. Otherwise commit the proposed optimizer state with the accepted parameters. Feasibility does not imply objective decrease or control the optimization gap.

The executed routine uses a ten-step power estimate to trigger spectral projection at radius 10 and clips biases at 10. It does not implement the class certificate or feasibility backtracking above. The runs are therefore diagnostic-only; observed maxima cannot replace domain suprema in Propositions 11 and 16.

D.4 Diagnostics and statistical reporting

For a paired input X'_s, the suffix $X_{s+1:h}$ is held fixed and the network is rerun. We record

$$I^Z_{h,s} = \frac{\|Z_h - Z_h^{(s)}\|_2}{\|X_s - X'_s\|_2}, \qquad I^\pi_{h,s} = \mathrm{TV}\{\pi_h(\cdot \mid X_{1:h}), \pi_h(\cdot \mid X^{(s)}_{1:h})\}.$$

Because later inputs are teacher-forced, these quantities match the perturbation in Proposition 1; they are blocked-feedback network interventions, not closed-loop causal effects.

In finite-state environments, exact filtering supplies b_h for evaluation. The audit records the mean, 95th percentile, and held-out maximum of $\mathrm{TV}(\hat{b}_h, b_h)$, together with δ_h and the pathwise right side of Theorem 5. For critic evaluation, a fixed history bank is paired with independent Monte Carlo continuations under the current policy. The audit separates Monte Carlo standard error from critic error. The reported regression in Table 8 uses log sensitivity summaries. A broader mixed regression would also relate critic error to sample size, $\widehat{K}^{Q,\mathrm{loc}}_{\mathrm{net}}$, horizon, environment, and seed; it is descriptive because Proposition 11 is an upper bound, not an exact regression law.

Primary baselines and complete ablations use 20 paired seeds; four E4 ablations have five complete seeds. Supplementary diagnostics state their counts separately. No completed ten-seed validation search is claimed. At each training checkpoint, policies are evaluated with 2,000 common-random-number episodes; the locked final checkpoint uses 10,000. Cumulative pseudo-regret sums the checkpoint-specific value estimates with their deployment weights m_k. If only a subset of epochs was evaluated, the interpolation or aggregation rule must be recovered from the evaluation code; sparse checkpoint sums alone do not define Regret_{N_K}. The reported intervals quantify between-seed variability; checkpoint-level Monte Carlo standard errors are not displayed and should be included in the archival logs. Intervals are 95% percentile bootstrap intervals over seeds using 10,000 resamples; paired contrasts resample paired seed differences. Correlations and regressions cluster repeated checkpoints by seed and environment.

E Synthetic Environments

Independent Rademacher distractors $N_h \in \{-1, +1\}^{d_N}$ are appended to observations. When an environment has no meaningful preterminal decision, a stage-dependent action mask forces a no-op, preventing actions from serving as an external memory channel.

E.1 Analytic LSTM cell audit

Use a scalar LSTM with constant input and output gates $I_h = O_h = 0.95$, constant forget gate $F_h = f \in \{0.05, 0.20, 0.50, 0.80\}$, and

$$G_h = \tanh(wX_h + uU_{h-1}), \quad w = 1, \quad u \in \{0, 0.5, 0.9, 1.1\}, \quad C_0 = U_0 = 0.$$

Paired inputs have $X_1 \in \{-0.1, +0.1\}$ and $X_h = 0$ thereafter, for $D \in \{2, 4, 8, 16, 32\}$ transitions. When $u = 0$, the direct retention route dominates; $u > 0$ creates a hidden-state-to-candidate path omitted by a forget-only score. Every derivative and finite intervention can be evaluated without fitting an RL policy.

E.2 Delayed recall

For delay D, set $H = D + 1$, draw $B \sim \text{Bernoulli}(1/2)$, and reveal $Q = B \oplus E$, where $E \sim \text{Bernoulli}(q)$. The observation is

$$Y_1 = (1, 0, Q, N_1), \qquad Y_h = (0, 0, 0, N_h) \ \ (1 < h < H), \qquad Y_H = (0, 1, 0, N_H).$$

All preterminal actions are forced no-ops and rewards are zero. At H, $A_H \in \{0, 1\}$ and $R_H = \mathbf{1}\{A_H = B\}$. The optimal value is $1 - q$. The prespecified primary setting is $(D, q, d_N) = (64, 0.1, 16)$. The delay grid is $D \in \{4, 8, 16, 32, 64, 128\}$, with one-factor robustness checks over $q \in \{0, 0.1, 0.2\}$ and $d_N \in \{0, 16, 64\}$.

E.3 Conditional recall

Fix $H = 65$. Draw $C \sim \text{Bernoulli}(p_L)$ and independent uniform bits B_L, B_S. The context is visible at every step, B_L only at $h = 1$, and B_S only at $h = H - d_S$; marker coordinates identify cue and query steps. The terminal target and reward are

$$T = CB_L + (1 - C)B_S, \qquad R_H = \mathbf{1}\{A_H = T\}.$$

Thus the same network needs a 64-transition memory when $C = 1$ and a short memory when $C = 0$. The primary setting is $(p_L, d_S, d_N) = (1/2, 4, 16)$; robustness varies $p_L \in \{1/4, 1/2, 3/4\}$ and $d_S \in \{2, 4, 8\}$.

E.4 Exactly filter-stable chain

Let $S_1 \sim \text{Bernoulli}(1/2)$ and

$$\mathbb{P}(S_{h+1} = s' \mid S_h = s) = \rho \mathbf{1}\{s' = s\} + \frac{1 - \rho}{2}. \tag{30}$$

At each step, $Y_h = S_h$ with probability $1 - \zeta$ and $Y_h = \perp$ with probability ζ. Preterminal actions are no-ops and the terminal reward is $\mathbf{1}\{A_H = S_H\}$. If $b_h = \mathbb{P}(S_h = 1 \mid H_h)$, then

$$\bar{b}_1 = \tfrac{1}{2}, \qquad \bar{b}_h = \tfrac{1}{2} + \rho(b_{h-1} - \tfrac{1}{2}), \qquad b_h = \begin{cases} 0, & Y_h = 0, \\ 1, & Y_h = 1, \\ \bar{b}_h, & Y_h = \perp. \end{cases}$$

A reveal has contraction coefficient zero and an erasure has coefficient ρ, so the uniform coefficient is exactly $\rho_B = \rho$. The optimal value is

$$V^\star = \frac{1}{2} + \frac{1 - \zeta}{2} \frac{1 - (\zeta\rho)^H}{1 - \zeta\rho}.$$

The primary setting is $(H, \rho, \zeta, d_N) = (32, 0.90, 0.75, 16)$, with one-factor grids

$$\rho \in \{0.20, 0.60, 0.90, 0.95\}, \quad \zeta \in \{0.25, 0.75, 0.95\}, \quad H \in \{8, 16, 32, 64\}.$$

For $H \le 12, d_N = 0$, the audit can enumerate the full history tree to avoid Monte Carlo error in action, occupancy, and value discrepancies.

E.5 Active sensing

A uniform bit B remains fixed for the episode. At $h < H$, $A_h \in \{\text{PROBE}, \text{SKIP}\}$. A probe makes the next observation $B \oplus E_{h+1}$, with $E_{h+1} \sim \text{Bernoulli}(q)$; a skip yields an erasure. The state-independent saving is

$$R_h = \frac{c}{H - 1} \mathbf{1}\{A_h = \text{SKIP}\}.$$

At H, the action predicts B and $R_H = \mathbf{1}\{A_H = B\}$. The exact log-odds belief and backward induction give $V^\star$. We use $H \in \{16, 32\}$, $q \in \{0.15, 0.25\}$, and $c \in \{0.05, 0.15\}$, with $(32, 0.25, 0.05)$ primary. Since probing changes future information, this family tests adaptive data collection but is not used to validate Assumption 4.

F Priority Experiments

The completed follow-ups in Section A.4 cover shorter delays, distractor removal, separate E2 context accuracies, learning curves, and same-update E4 exact-belief controls with probe frequencies. They do not demonstrate a reliable regret benefit from including influence regularization. The separate study in Section A.5 now confirms a benefit when old information is unnecessary; the partially observed pilot conditions remain exploratory.

Further experiments should compare the implemented terminal-reward block extension with GAMPO-P-R's separately weighted terminal head under matched budgets, reporting head sharing, λ_{T}, and terminal reward prediction error. Decision-stage averaging is already implemented; any all-stage comparison should be a separately labeled ablation. A supervised terminal-cue control and an oracle action-value critic could separate remaining representation and optimization difficulties more directly than the exact-belief inputs, which still require learned value heads.

Across λ_{Γ}, report cue and nuisance influence, paired finite interventions, regret, and final value. Compare finite effects with a certified tail only when a global certificate is available. Independently validated recurrent RL implementations and standard-suite evaluations remain necessary before broader performance claims. No favorable outcome of these unperformed checks is assumed.

Reproducibility statement

The appendix supplies proofs, objectives, environment definitions, and the available experimental record. The deterministic E0 audit was recomputed with explicit absolute and relative tolerances. Primary RL tables are reaggregated from archived seed-level measurements, and supplementary controls add actual training runs; neither evaluates GAMPO-P-R. Section D specifies architectures, executed losses, optimization, and evaluation.

AI use statement

Generative AI assisted language editing, literature checks, mathematical derivations, interpretation of the supplied results, and methodological revision. AI-assisted computation recomputed the deterministic E0 grid. Supplementary RL diagnostics were implemented and executed on HPCC; their measurements come from 170 completed runs and retained records (85 initial diagnostics, 45 pilot runs, and 40 confirmation runs). The authors are responsible for checking the revised proofs, implementation, citations, and numerical records before submission.