NOT EVERY OBSERVATION SHOULD UPDATE BELIEF EQUALLY

Anonymous authors
Paper under double-blind review

ABSTRACT

Partially observable reinforcement learning requires agents to maintain an internal state from incomplete observations. Recurrent agents can modulate how incoming observations affect their internal state through learned gates, yet the strength of each revision is typically learned implicitly through end-to-end control. Observation surprise and decision relevance are distinct: a surprising observation may be irrelevant, while a plausible cue may alter the preferred action. We introduce BELIEFGATE, which makes revision strength an explicit decision variable supervised by downstream control utility. During training, paired interventions compare candidate revision strengths from the same simulator state under matched future randomness, and penalized downstream utility selects the target for a continuous deployment-time gate. We evaluate matched-surprise and same-token diagnostics together with long-memory and observation-shift benchmarks. BELIEFGATE achieves the highest overall and unseen-shift aggregates, while specialized memory baselines remain stronger on some memory-centric tasks. Under matched surprise, it achieves the strongest non-oracle separation between relevant and irrelevant observations and exhibits the smallest relative degradation under observation shifts. Representation probes reveal more decision-relevant content and less distractor retention. These results support treating observation assimilation as a decision-relevant control problem rather than a generic filtering operation.

Keywords: partially observable reinforcement learning; belief representation; observation assimilation; recurrent agents; decision relevance.

1 INTRODUCTION

Partial observability requires an agent to compress history into an internal state because the current observation alone is insufficient for control. This challenge spans classical POMDPs, long-memory benchmarks, and modern settings with partially exposed reward-relevant state (Kaelbling et al., 1998; Morad et al., 2023; Lang et al., 2024; Kausik et al., 2025; Huang et al., 2025). The practical challenge is that the agent must decide how strongly each new observation should revise the state used for decision making. Effective control therefore requires both retaining relevant history and regulating how incoming evidence revises that history-dependent state.

Recurrent agents already implement selective state updates through gated recurrence, replay-based recurrent training, and specialized memory architectures (Hochreiter and Schmidhuber, 1997; Kapturowski et al., 2019; Ni et al., 2022; Morad et al., 2024; Le et al., 2025; Eberhard et al., 2025). Yet the strength of each revision is typically learned implicitly through return optimization. Their objectives therefore provide no direct supervision for whether a particular observation-induced revision improves future control.

Observation surprise and decision relevance are not equivalent. A large innovation can reflect a nuisance change, while a predictable cue can still alter the preferred action. Two observations with matched surprise can therefore warrant very different revisions. Robustness, representation-learning, and value-of-information objectives address adjacent questions (Nakanishi et al., 2025; Liang et al., 2025; Laouar et al., 2026), but do not directly specify how strongly an already received observation should revise the learned belief representation. The missing supervision target is therefore not observation novelty, but the downstream control value of the resulting belief revision. We define

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

the preferred revision counterfactually by comparing future control outcomes when the same prior assimilates the same observation at different strengths.

We introduce BELIEFGATE, which separates a recurrent belief prior from an observation-conditioned proposal and controls their interpolation with an explicit scalar $\alpha_t \in [0, 1]$. During training, matched counterfactual branches compare candidate revision strengths from the same simulator state, and penalized downstream utility provides direct supervision for the gate. At deployment, the branches are removed; the learned gate uses the current belief prior, observation embedding, and innovation residual to perform a single state update. Across diagnostic POMDPs, memory benchmarks, and observation shifts, the learned gate better separates decision-relevant from irrelevant observations, while the resulting policy shows smaller degradation under observation shifts; specialized memory methods remain stronger on some memory-centric settings.

Our contributions are threefold:

(1) **Problem formulation.** We formulate observation assimilation as a decision-relevant control problem by making revision strength explicit and supervising it according to downstream control utility rather than observation novelty.
(2) **Learning principle.** We introduce paired counterfactual utility supervision: sparse interventions compare alternative revision strengths from the same simulator state under matched future randomness, and the preferred revision strength is distilled into a continuous single-pass deployment gate.
(3) **Mechanistic validation.** We develop matched-surprise and same-token diagnostics that isolate decision relevance from observation surprise and cue identity, and combine them with utility-target audits, held-out shifts, and representation probes to test whether the learned gate tracks downstream decision consequences rather than superficial observation statistics.

2 RELATED WORK

Memory and Belief Representations. Classical POMDP control reasons over beliefs (Kaelbling et al., 1998), whereas deep RL learns internal state with recurrent and sequence models, including replay-based recurrent agents (Kapturowski et al., 2019), transformer-style sequence models (Parisotto et al., 2020), and specialized memory mechanisms (Morad et al., 2024). Predictive and control-oriented representations structure latent state for decision making (Littman et al., 2001; Gelada et al., 2019; Misra et al., 2024), while world-model agents learn compact latent dynamics for control and memory (Hafner et al., 2019; 2020; Samsami et al., 2024). Across these approaches, recurrent and memory design primarily determines what information is represented, retained, and propagated, with selective recurrence already providing implicit control over information flow. BELIEFGATE instead focuses on revision supervision: it exposes the magnitude of the revision induced by the current observation as an explicit variable supervised by downstream control utility.

Observation Robustness, Relevance, and Selective Processing. Robust RL perturbs observations or optimizes against adversarial inputs (Gu et al., 2025; Nakanishi et al., 2025); reconstruction and fusion methods seek task-relevant latent structure (Wang et al., 2025; Liang et al., 2025); and distractor-focused work studies nuisance-robust representations (Nikulin et al., 2025). These lines primarily address robustness to corruption, recovery of task-relevant structure, or suppression of nuisance variation. Value-of-information planning asks whether additional information processing or belief-space reasoning is worth its computational cost (Laouar et al., 2026). BELIEFGATE instead models the degree to which an already received observation should revise the current belief representation. Its supervision signal is the downstream control consequence of that revision, rather than input quality, reconstruction fidelity, or planning allocation.

Training-Time Supervision and Counterfactual Signals. Privileged information can improve partially observable learning during training (Cai et al., 2024; Huang et al., 2025), hindsight observations can reduce sample complexity (Guo et al., 2024), and teacher–student objectives provide additional training signals (Messikommer et al., 2025). Such supervision supplies extra state, observations, or teacher targets during learning. BELIEFGATE instead constructs its supervision target by intervening on the agent's own revision strength: paired branches begin from the same simulator state, vary the current revision, and compare downstream control utility under matched future randomness. The

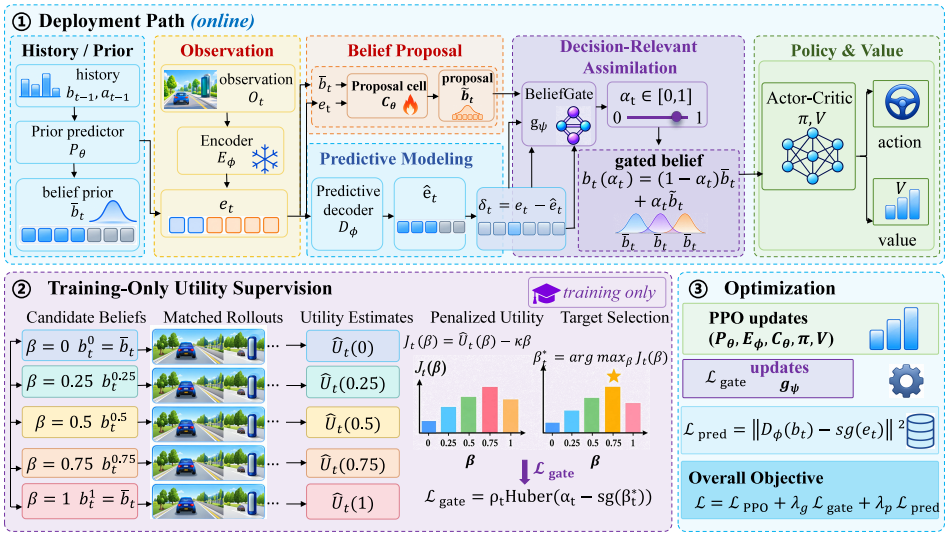


Figure 1: BeliefGate framework. The deployment path combines a recurrent prior and an observation-conditioned proposal through a learned assimilation gate. During training, matched rollouts compare candidate revision strengths and supervise the gate using penalized downstream utility.

selected revision strength is then distilled into the deployment gate, while the extra branches disappear at test time.

3 METHOD

A recurrent update often entangles two operations: constructing an observation-conditioned proposal and deciding how strongly it should revise the current belief representation. BELIEFGATE separates them. Given a learned belief prior $\bar{b}_t$, the current observation produces a proposal $\tilde{b}_t$, and a scalar gate α_t sets the interpolation strength. Sparse training-only branches intervene on α_t to estimate the downstream consequence of each revision. Figure 1 separates the deployment path from the paired supervision path.

We consider a POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, R, \gamma)$ with latent state $s_t \in \mathcal{S}$, action $a_t \in \mathcal{A}$, observation $o_t \in \mathcal{O}$, transition kernel T, observation kernel Z, reward R, and discount $\gamma \in (0, 1)$. The agent maintains $b_t \in \mathbb{R}^d$ as a learned belief representation. Here, *belief* denotes the learned recurrent representation used for control, not necessarily an exact Bayesian posterior. From the previous belief and action we form

$$\bar{b}_t = P_\theta(b_{t-1}, \phi_a(a_{t-1})), \qquad e_t = E_\theta(o_t), \tag{1}$$

where P_θ is an action-conditioned recurrent predictor, ϕ_a embeds the previous action, and E_θ is an observation encoder.

3.1 DECISION-RELEVANT ASSIMILATION

The gate predicts the current revision strength from the belief prior, observation embedding, and innovation residual, while its training target is defined by downstream control utility.

The assimilation proposal is

$$\tilde{b}_t = C_\theta(\bar{b}_t, e_t), \tag{2}$$

where C_θ is an update cell from the same recurrent family as the baseline. A lightweight decoder predicts the incoming embedding from the prior, $\hat{e}_t = D_\theta(\bar{b}_t)$, and the innovation residual is $\delta_t = e_t - \hat{e}_t$. The gate outputs

$$\alpha_t = \sigma\big(g_\psi(\bar{b}_t, e_t, \delta_t)\big), \qquad b_t(\alpha_t) = (1 - \alpha_t)\bar{b}_t + \alpha_t\tilde{b}_t. \tag{3}$$

Here g_ψ is a two-layer MLP and σ is the logistic function. Equation (3) provides an explicit axis for intervention; Section 4.2 evaluates the fidelity of the resulting utility targets.

3.2 Paired Utility Supervision

Policy gradients assign delayed credit jointly across hidden-state operations. We instead create sparse counterfactual interventions at selected states to obtain a direct revision target. Let $\mathcal{B} = \{0, .25, .5, .75, 1\}$ be candidate revision strengths. At a probe time t, branch $\beta \in \mathcal{B}$ starts from

$$b_t^\beta = (1-\beta)\bar{b}_t + \beta\tilde{b}_t. \tag{4}$$

After this one-step intervention, all branches return to the current target copy of BeliefGate: at $t+1, \ldots, t+H$, observations are assimilated using $\alpha_{t+j}^- = \sigma\big(g_{\psi^-}(\bar{b}_{t+j}, e_{t+j}, \delta_{t+j})\big)$ rather than keeping β fixed. Thus $\hat{U}_t(\beta)$ measures the consequence of changing only the current revision.

For each of M paired random seeds, branches share the same simulator snapshot and exogenous random variates. When actions are sampled stochastically, branches also reuse the same base random numbers so that policy randomness is matched across rollouts. The branch score is

$$\hat{U}_t(\beta) = \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{j=0}^{H-1}\gamma^j r_{t+j}^{m,\beta} + \gamma^H V_{\omega^-}(b_{t+H}^{m,\beta})\right], \tag{5}$$

where V_{ω^-} and g_{ψ^-} are slowly updated target networks. Branch rollouts generate supervision labels only, so gradients do not pass through Eq. (5). Branch utilities use the native within-task return scale without additional normalization.

To avoid averaging across potentially multimodal utility profiles, we select the supervision target by penalized utility

$$J_t(\beta) = \hat{U}_t(\beta) - \kappa \cdot \beta, \qquad \beta_t^\star = \arg\max_{\beta\in\mathcal{B}} J_t(\beta), \tag{6}$$

where $\kappa \geq 0$ penalizes stronger current revisions. This creates a conservative default: a stronger revision must provide enough additional utility to justify placing more weight on the observation-conditioned proposal. Ties within numerical tolerance are resolved toward the smallest β. The scalar gate remains continuous; only its supervision target is selected from $\mathcal{B}$. We weight informative probes by

$$\rho_t = \epsilon + (1-\epsilon)\tanh\left(\frac{\max_\beta \hat{U}_t(\beta) - \min_\beta \hat{U}_t(\beta)}{c}\right), \tag{7}$$

with $c > 0$ a utility-scale constant and $\epsilon > 0$ a minimum weight. When candidate utilities are nearly indistinguishable, the preferred revision is weakly identified; ρ_t therefore downweights such probes. The gate loss is

$$\mathcal{L}_{\text{gate}} = \mathbb{E}_t\left[\rho_t \ \text{Huber}(\alpha_t - \text{sg}(\beta_t^\star))\right]. \tag{8}$$

3.3 Optimization and Deployment

We optimize the policy and gate through separate gradient paths. PPO (Schulman et al., 2017) updates the encoder, predictor, proposal cell, actor, and critic while treating α_t as a stop-gradient scalar with respect to ψ. Conversely, $\mathcal{L}_{\text{gate}}$ updates only ψ, with $(\bar{b}_t, e_t, \delta_t)$ detached for this loss. The predictive auxiliary term

$$\mathcal{L}_{\text{pred}} = \left\|D_\theta(\bar{b}_t) - \text{sg}(e_t)\right\|_2^2 \tag{9}$$

updates the predictor/decoder path and is separately ablated. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{PPO}} + \lambda_g\mathcal{L}_{\text{gate}} + \lambda_p\mathcal{L}_{\text{pred}}. \tag{10}$$

At deployment, Eqs. (1)–(3) run once per step; paired branches, target critics, and Eq. (5) are removed. Branch simulator steps are counted separately from policy-training interactions. Table 1 reports parameters, training overhead, and same-device inference latency; Table 12 provides the detailed computational profile for matched recurrent and gate baselines.

4 Experiments

We ask whether utility supervision makes assimilation track decision relevance rather than observation novelty. We evaluate this through aggregate performance, matched-surprise diagnostics, utility-target validity, observation shifts, representation probes, and ablations. Detailed breakdowns appear in the appendix.

Table 1: Overall performance, hard-case transfer, and deployment cost.

Panel A: aggregate benchmark performance and deployment cost

Method	Controlled IQM↑	POPGym IQM↑	Delay/Memory IQM↑	Unseen-shift IQM↑	Overall IQM↑	Params (M) ↓	Train overhead ↓	Infer. ms ↓
GRU-PPO	0.668	0.594	0.550	0.477	0.582	1.52	1.00×	0.31
R2D2-style	0.700	0.649	0.626	0.532	0.640	1.55	1.19×	0.36
Memoroid	0.729	0.756	0.686	0.580	0.703	1.49	1.11×	0.34
Stable Hadamard Memory	0.760	0.824	0.748	0.646	0.758	1.54	1.31×	0.38
Memory Traces	0.736	0.788	0.839	0.648	0.757	1.47	1.08×	0.29
Policy-POMDP (2025)	0.762	0.711	0.741	0.670	0.732	1.50	1.17×	0.33
Adv.-Obs. AC	0.689	0.609	0.583	0.579	0.630	1.51	1.34×	0.41
BeliefGate	0.884	0.814	0.833	0.778	0.841	1.53	2.36×	0.35

Panel B: hard-case slices (normalized return unless noted)

Method	Same-token	Sparse cue	Long delay	Heavy alias	Burst + delay	Missing + distractor	Hard-case IQM↑	Failure rate↓
GRU-PPO	0.548	0.603	0.458	0.519	0.398	0.468	0.497	0.390
R2D2-style	0.613	0.658	0.582	0.641	0.501	0.526	0.591	0.350
Memoroid	0.651	0.724	0.673	0.662	0.557	0.604	0.648	0.240
Stable Hadamard Memory	0.692	0.812	0.742	0.731	0.632	0.667	0.713	0.220
Memory Traces	0.665	0.768	0.851	0.681	0.697	0.641	0.724	0.270
Policy-POMDP (2025)	0.728	0.686	0.714	0.764	0.584	0.656	0.701	0.310
Adv.-Obs. AC	0.629	0.617	0.561	0.603	0.736	0.856	0.657	0.180
BeliefGate	0.884	0.803	0.829	0.791	0.796	0.842	0.823	0.120

4.1 Experimental Protocol and Baselines

We use three non-overlapping benchmark groups. **Controlled Diagnostics** include CueSwitch, AliasChain, KeyDoor, and DistractorControl, with independent factors for corruption, relevance, delay, aliasing, and distractor strength. A same-token variant presents the identical observation under histories in which it either changes or preserves the optimal action. **POPGym** provides standardized partially observable memory tasks (Morad et al., 2023). **Delay/Memory-Control** uses distinct delay-oriented settings motivated by Wang et al. (2024) and Anokhin et al. (2025). For robustness evaluation, we apply cue swaps, missing frames, burst noise, aliasing, delays, and distractors where appropriate. Development uses five seeds; final results are aggregated over ten fixed seeds with paired evaluation. Cross-task summaries use IQM of task-normalized return; gate, ablation, and sensitivity panels report their column-specific aggregates over the same ten final seeds.

The unified comparison includes GRU-PPO (Ni et al., 2022), R2D2-style recurrence (Kapturowski et al., 2019), Memoroid (Morad et al., 2024), Stable Hadamard Memory (Le et al., 2025), Memory Traces (Eberhard et al., 2025), a scalable policy-based POMDP method (Anjarlekar et al., 2025), and adversarial-observation actor–critic (Nakanishi et al., 2025). Recurrent and memory baselines use a common PPO outer protocol; "R2D2-style" refers to the recurrent architecture rather than the original distributed R2D2 training algorithm. Comparable recurrent models are capacity-matched before training by adjusting encoder and recurrent widths to keep total trainable parameters within 5% of BeliefGate while preserving each architecture. Diagnostic gate variants—PPO-only, surprise-only, fixed-α, and oracle utility—share the same explicit-assimilation interface. Setting-specific methods with different information or interaction assumptions are included as contextual references (Liang et al., 2025; Kang et al., 2025; Mahjoub et al., 2025; Kontogiannis et al., 2025; Zhou et al., 2025).

Across the aggregate benchmarks, BeliefGate leads on Controlled, unseen-shift, and Overall IQM, while Stable Hadamard Memory is stronger on POPGym and Memory Traces on Delay/Memory (Table 1). On the hard-case slices, it achieves the best aggregate IQM and lowest failure rate, while specialized baselines remain stronger on sparse cues, long delay, and missing-plus-distractor settings.

For each task τ, episode return is normalized as

$$\widetilde{R}_e^{(\tau)} = \frac{R_e^{(\tau)} - R_{\text{rand}}^{(\tau)}}{R_{\text{ref}}^{(\tau)} - R_{\text{rand}}^{(\tau)}}.$$

Here $R_{\text{rand}}^{(\tau)}$ is the return of a uniform-random action policy, and $R_{\text{ref}}^{(\tau)}$ is obtained from a separately trained fully observable, privileged-state reference policy for the same task. The task-specific anchors

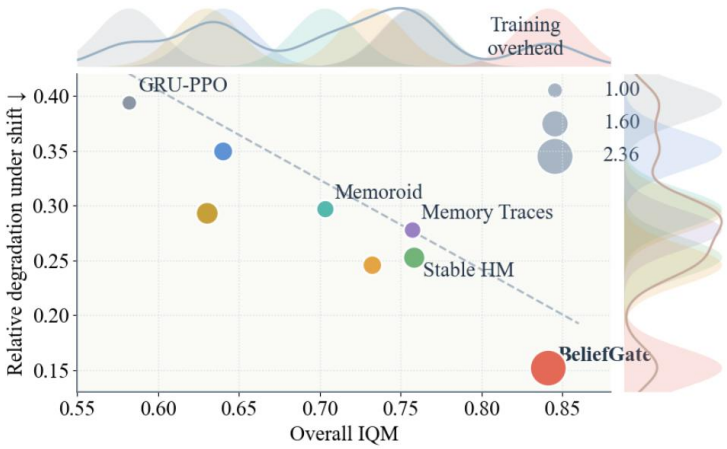


Figure 2: Overall performance and robustness. Each marker denotes one method; the horizontal axis reports Overall IQM, the vertical axis reports relative drop across the base observation-shift suite, and marker size indicates training overhead.

are estimated during development and frozen before final evaluation; normalized returns are not clipped. Failure is defined as $\widetilde{R}_e^{(\tau)} < 0.25$, with the threshold fixed before final-seed evaluation.

Figure 2 summarizes overall IQM, relative shift degradation, and training overhead. Paired branches increase training cost, whereas deployment latency remains in the range of the recurrent baselines. Per-task and detailed cost results appear in Tables 13 and 12.

4.2 Does the Gate Track Decision Relevance?

Our central claim is that assimilation should reflect decision relevance rather than observation surprise. We test this claim with a common paired set in which reference surprise is matched while decision relevance changes. The set is constructed using a reference encoder–predictor trained on the development split and frozen before final evaluation. Candidate states are stratified by task, corruption family/severity, cue-to-decision delay, token identity, and token-frequency bin before matching the reference innovation $\delta_t^{\mathrm{ref}} = e_t^{\mathrm{ref}} - \hat{e}_t^{\mathrm{ref}}$. Within each stratum, pairs minimize

$$\Delta_{\mathrm{surprise}} = \left| \|\delta_t^{\mathrm{ref},(r)}\|_2 - \|\delta_t^{\mathrm{ref},(i)}\|_2 \right|. \tag{11}$$

Decision-relevance labels are provided by the controlled-task oracle and are independent of the learned branch target. Let h_t^- denote the history before receiving o_t. The oracle defines

$$y_t^{\mathrm{rel}} = \mathbb{I}\left[\arg\max_a Q^{\mathrm{orc}}(h_t^-, o_t, a) \neq \arg\max_a Q^{\mathrm{orc}}(h_t^-, \varnothing, a) \right], \tag{12}$$

Here $\varnothing$ withholds the current cue under the matched counterfactual. The same-token variant fixes the observation while changing its history; all methods are evaluated on the same paired histories, with nuisance factors balanced before surprise matching.

Under matched surprise, BeliefGate assigns higher revision strength to oracle-relevant observations and lower revision strength to irrelevant ones, whereas the surprise-only gate shows little separation. The fixed gate attains perfect decision-flip recall, but because it applies the same revision strength everywhere, its stable-case specificity is zero. For utility-target validity, $H = 8$ gives the best argmax agreement, MC regret, action disagreement, and single-mode rate. Although $H = 16$ slightly improves rank correlation, it does not improve these selection-oriented metrics, motivating the default $H = 8$. BeliefGate maintains this separation across near-, mid-, and far-delay tests and in the same-token setting. Recall and specificity use a single τ_α selected on development seeds by balanced accuracy against y_t^{rel} and frozen for all final-seed slices.

4.3 Robustness, Generalization, and Representation

Training uses cue swap, missing-observation, burst-noise, and distractor interventions within a bounded corruption range; aliasing and delay are held out as unseen intervention families. Evaluation

Table 2: Core mechanism validation: matched-surprise behavior, utility-target validity, and context-dependent relevance.

Panel A: matched-surprise gate behavior						
Gate	Surprise gap↓	$\alpha_r \uparrow$	$\alpha_i \downarrow$	$\Delta\alpha \uparrow$	Rel. AUROC↑	Matched-set return↑
PPO-only gate	0.019	0.710	0.460	0.250	0.690	0.782
Surprise-only gate	0.019	0.720	0.650	0.070	0.550	0.816
Fixed $\alpha = .5$	0.019	0.500	0.500	0.000	0.500	0.803
Oracle utility gate	0.019	0.910	0.160	0.750	0.985	0.934
BeliefGate	0.019	0.840	0.340	0.500	0.910	0.892

Panel B: bootstrap utility-target validity (all metrics against long-MC reference)						
Target estimator	Rank corr.↑	Argmax agr.↑	MC regret↓	Action disagree↓	Single-mode % ↑	Local maxima↓
Critic-only ($H = 0$)	0.612	0.548	0.086	0.327	57.4	1.72
Bootstrap $H = 4$	0.773	0.704	0.049	0.214	69.8	1.41
Bootstrap $H = 8$	0.861	0.824	0.026	0.118	82.1	1.19
Bootstrap $H = 16$	0.878	0.806	0.031	0.139	77.3	1.28

Panel C: context and temporal stress tests for decision relevance						
Gate	Near-cue $\Delta\alpha \uparrow$	Mid-delay $\Delta\alpha \uparrow$	Far-delay $\Delta\alpha \uparrow$	Same-token Rel. AUROC↑	Decision-flip recall↑	Stable-case specificity↑
PPO-only gate	0.286	0.211	0.137	0.671	0.722	0.658
Surprise-only gate	0.153	0.069	0.028	0.518	0.602	0.494
Fixed $\alpha = .5$	0.000	0.000	0.000	0.500	1.000	0.000
Oracle utility gate	0.792	0.744	0.689	0.981	0.956	0.943
BeliefGate	0.574	0.516	0.438	0.903	0.889	0.857

Table 3: Robustness and generalization across shift type, composition, and severity.

Panel A: shift-family generalization						
Method	Clean↑	Seen Shift IQM↑	Unseen Shift IQM↑	Unseen Combo IQM↑	Worst Shift↑	Rel. Drop↓
GRU-PPO	0.756	0.582	0.477	0.412	0.458	0.394
R2D2-style	0.788	0.627	0.532	0.478	0.512	0.350
Memoroid	0.812	0.671	0.580	0.526	0.571	0.297
Stable Hadamard Memory	0.846	0.711	0.646	0.586	0.632	0.253
Memory Traces	0.835	0.684	0.648	0.574	0.603	0.278
Policy-POMDP (2025)	0.824	0.675	0.670	0.591	0.621	0.246
Adv.-Obs. AC	0.782	0.733	0.579	0.641	0.553	0.293
BeliefGate	0.906	0.825	0.778	0.748	0.768	0.152

Panel B: severity ladder and compositional stress						
Method	Mild OOD ↑	Medium OOD ↑	Severe OOD ↑	2-shift combo↑	3-shift combo↑	Severity slope ↓
GRU-PPO	0.668	0.571	0.413	0.492	0.361	0.128
R2D2-style	0.719	0.624	0.476	0.548	0.414	0.122
Memoroid	0.756	0.687	0.559	0.608	0.469	0.099
Stable Hadamard Memory	0.808	0.746	0.632	0.671	0.531	0.088
Memory Traces	0.792	0.724	0.611	0.653	0.512	0.091
Policy-POMDP (2025)	0.781	0.704	0.576	0.621	0.493	0.103
Adv.-Obs. AC	0.804	0.754	0.681	0.713	0.646	0.062
BeliefGate	0.874	0.832	0.761	0.801	0.724	0.057

additionally includes unseen combinations and stronger corruption. Frozen-representation probes use common held-out replay prefixes and assess latent-state recovery, next-critical-event prediction, distractor identity, and branch-point action.

BELIEFGATE degrades more slowly under distribution shift, with the lowest relative drop and severity slope in Table 3. It also achieves the highest normalized return under unseen combinations and severe corruption; Adv.-Obs. AC is the closest baseline in the severe and multi-shift settings. Relative drop is computed as $(R_{\text{clean}} - R_{\text{worst}})/R_{\text{clean}}$, where R_{worst} is the minimum normalized return across the six base single-shift interventions. Severity slope is the absolute least-squares slope over three held-out OOD intensities. The representation probes show a complementary pattern: BELIEFGATE yields higher critical-event AUROC and branch-action accuracy while retaining less information about distractor identity, without uniformly maximizing generic latent-state recovery (Figure 3; Table 14). This suggests that utility-supervised assimilation affects which information persists in the belief representation, rather than only the instantaneous gate magnitude. Across all six base shift

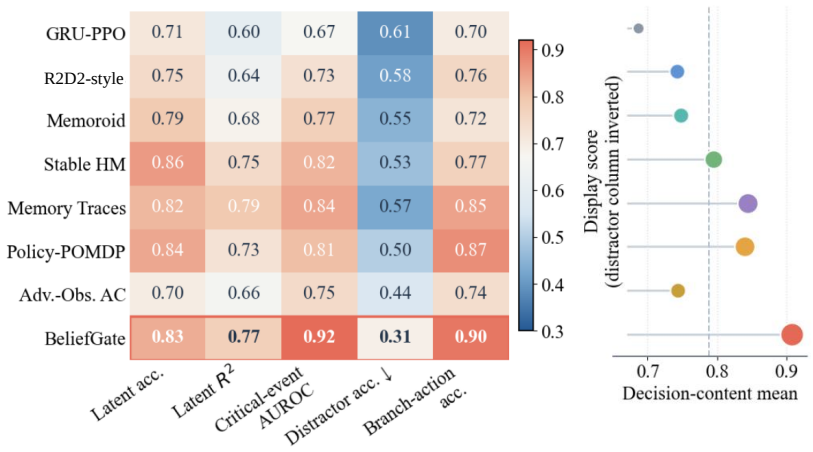


Figure 3: Representation analysis. The heatmap summarizes latent-state probing across five criteria: latent-state accuracy, latent-state R^2, critical-event AUROC, distractor accuracy, and branch-action accuracy. The right panel reports the *Decision-content mean*, averaged over critical-event AUROC and branch-action accuracy. Distractor accuracy is reported raw and is lower-is-better.

Table 4: Ablation and key sensitivity of decision-relevant assimilation.

Panel A: component ablations Variant	Return↑	$\Delta\alpha$ ↑	U-regret↓	Worst shift↑	State probe↑
Full BeliefGate	0.861	0.487	0.028	0.768	0.824
PPO-only gate	0.792	0.218	0.057	0.612	0.756
Surprise-only gate	0.774	0.074	0.066	0.548	0.731
Fixed α	0.803	0.000	0.082	0.571	0.719
No prediction loss	0.836	0.451	0.032	0.718	0.655
No innovation δ_t	0.821	0.337	0.043	0.697	0.812
No revision penalty κ	0.848	0.536	0.046	0.739	0.827
Uniform probe weighting	0.829	0.421	0.040	0.712	0.804
Binary-target gate	0.832	0.558	0.042	0.686	0.816

Panel B: key compute–quality sensitivity slices (one factor changed from default) Setting	Return↑	$\Delta\alpha$ ↑	U-regret↓	Worst shift↑	State probe↑	Branch sim. ratio↓
Default (H=8, p=0.05, M=2, $\lvert\mathcal{B}\rvert$=5)	0.861	0.487	0.028	0.768	0.824	4.00
Short horizon (H=4)	0.842	0.454	0.040	0.743	0.812	2.00
Long horizon (H=16)	0.856	0.501	0.023	0.754	0.827	8.00
Sparse probing (p=0.02)	0.846	0.468	0.036	0.749	0.817	1.60
Dense probing (p=0.10)	0.864	0.493	0.027	0.771	0.825	8.00
Coarse candidates ($\lvert\mathcal{B}\rvert$=3)	0.848	0.459	0.037	0.741	0.819	2.40
Dense candidates ($\lvert\mathcal{B}\rvert$=9)	0.859	0.506	0.025	0.763	0.823	7.20

families, the gain over the strongest competing method remains positive rather than being driven by a single corruption type (Figure 4; Table 11).

4.4 Ablations and Key Sensitivities

The ablations isolate utility supervision (PPO-only and surprise-only gates), update design (fixed and binary-target gates; removal of δ_t), and training terms (removal of $\mathcal{L}_{\text{pred}}$, κ, or ρ_t). State probes use common held-out replay prefixes.

Among the component ablations, the full method gives the highest return and worst-shift score. Removing κ or using binary-target supervision increases $\Delta\alpha$ while reducing return and increasing utility regret, indicating that larger gate separation alone is insufficient. Removing the prediction loss produces the largest drop in the state-probe score. In Panel B, a longer horizon lowers utility regret without improving return, while denser probing gives only a small gain at higher branch cost. The

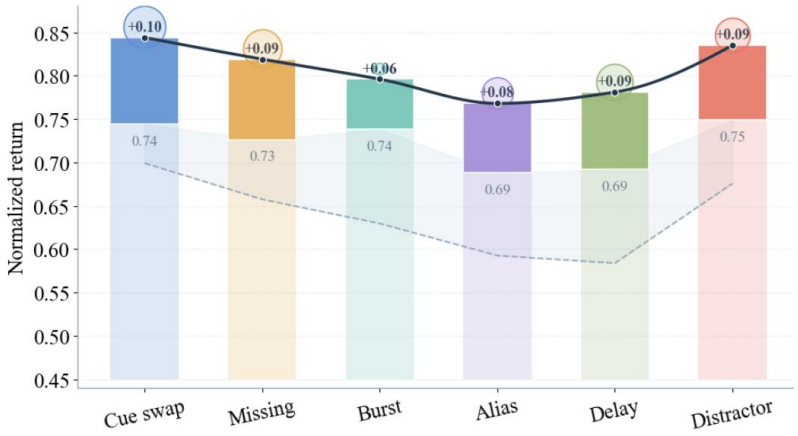


Figure 4: Shift-wise robustness gain. For each base observation shift, the pale bar shows the strongest non-BeliefGate baseline, while the dashed curve shows the mean non-BeliefGate baseline. The colored segment and bubble show BeliefGate's gain over the strongest baseline, while the solid curve shows its absolute return.

branch-simulation ratio is the number of additional paired simulator steps divided by base on-policy training steps; under the default setting it equals $p_{\text{probe}} M|\mathcal{B}|H = 4.0$.

5 Conclusion and Future Work

We studied observation assimilation as a decision-relevant control choice rather than a generic filtering operation. BeliefGate uses matched training-time counterfactuals to supervise how strongly each observation revises the learned belief representation while retaining single-pass deployment. Across matched-surprise diagnostics and partially observed control benchmarks, the results indicate that explicit utility supervision can produce belief updates that respond more selectively to decision-relevant evidence than to merely surprising input and improve robustness to observation shifts. The fact that specialized memory baselines remain stronger on some memory-centric tasks suggests that revision supervision and memory capacity address complementary aspects of partial observability. A natural next step is to replace finite candidate interventions with continuous utility modeling and extend decision-relevant assimilation to structured belief updates and model misspecification.

AI USE STATEMENT

Large language models were used for literature organization, language editing, LaTeX drafting, and consistency checks.

REPRODUCIBILITY STATEMENT

Section 3 specifies the BeliefGate method and training procedure, while Section 4 describes the evaluation protocol and reporting settings. Appendix A provides the main experimental configurations, hyperparameters, intervention definitions, baseline matching rules, and representation probes. We additionally provide anonymized supplementary code implementing the core BeliefGate mechanism and key training semantics.

REFERENCES

Ameya Anjarlekar, S. Rasoul Etesami, and R. Srikant. Scalable policy-based RL algorithms for POMDPs. In *Advances in Neural Information Processing Systems*, 2025.

Ivan Anokhin, Rishav Rishav, Matthew Riemer, Stephen Chung, Irina Rish, and Samira Ebrahimi Kahou. Handling delay in real-time reinforcement learning. In *International Conference on Learning Representations*, 2025.

Roman Belaire, Arunesh Sinha, and Pradeep Varakantham. On minimizing adversarial counterfactual error in adversarial reinforcement learning. In *International Conference on Learning Representations*, 2025.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. In *Advances in Neural Information Processing Systems*, 2024.

Onno Eberhard, Michael Muehlebach, and Claire Vernade. Partially observable reinforcement learning with memory traces. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Shangding Gu, Laixi Shi, Muning Wen, Ming Jin, Eric Mazumdar, Yuejie Chi, Adam Wierman, and Costas Spanos. Robust Gymnasium: A unified modular benchmark for robust reinforcement learning. In *International Conference on Learning Representations*, 2025.

Jiacheng Guo, Minshuo Chen, Huan Wang, Caiming Xiong, Mengdi Wang, and Yu Bai. Sample-efficient learning of POMDPs with multiple observations in hindsight. In *International Conference on Learning Representations*, 2024.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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

Dongchi Huang, Jiaqi Wang, Yang Li, Chunhe Xia, Tianle Zhang, and Kaige Zhang. PIGDreamer: Privileged information guided world models for safe partially observable reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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.

Sehyeok Kang, Yongsik Lee, Gahee Kim, Song Chong, and Se-Young Yun. MA^{2}E: Addressing partial observability in multi-agent reinforcement learning with masked auto-encoder. In *International Conference on Learning Representations*, 2025.

Steven Kapturowski, Georg Ostrovski, John Quan, Rémi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *International Conference on Learning Representations*, 2019.

Chinmaya Kausik, Mirco Mutti, Aldo Pacchiano, and Ambuj Tewari. A theoretical framework for partially-observed reward states in RLHF. In *International Conference on Learning Representations*, 2025.

Andreas Kontogiannis, Konstantinos Papathanasiou, Yi Shen, Giorgos Stamou, Michael M. Zavlanos, and George Vouros. Enhancing cooperative multi-agent reinforcement learning with state modelling and adversarial exploration. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, and Scott Emmons. When your AIs deceive you: Challenges of partial observability in reinforcement learning from human feedback. In *Advances in Neural Information Processing Systems*, 2024.

Zakariya Laouar, Qi Heng Ho, and Zachary Sunberg. Leveraging the value of information in POMDP planning. In *International Conference on Automated Planning and Scheduling*, 2026.

Hung Le, Dung Nguyen, Kien Do, Sunil Gupta, and Svetha Venkatesh. Stable Hadamard Memory: Revitalizing memory-augmented agents for reinforcement learning. In *International Conference on Learning Representations*, 2025.

Qifan Liang, Yixiang Shan, Haipeng Liu, Zhengbang Zhu, Ting Long, Weinan Zhang, and Yuan Tian. Reconstruction-guided policy: Enhancing decision-making through agent-wise state consistency. In *International Conference on Learning Representations*, 2025.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, 2001.

Omayma Mahjoub, Sasha Abramowitz, Ruan John De Kock, Wiem Khlifi, et al. Sable: A performant, efficient and scalable sequence model for multi-agent reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Nico Messikommer, Jiaxu Xing, Elie Aljalbout, and Davide Scaramuzza. Student-informed teacher training. In *International Conference on Learning Representations*, 2025.

Dipendra Kumar Misra, Akanksha Saran, Tengyang Xie, Alex Lamb, and John Langford. Towards principled representation learning from videos for reinforcement learning. In *International Conference on Learning Representations*, 2024.

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

Steven Morad, Chris Lu, Ryan Kortvelesy, Stephan Liwicki, Jakob Foerster, and Amanda Prorok. Recurrent reinforcement learning with memoroids. In *Advances in Neural Information Processing Systems*, 2024.

Kosuke Nakanishi, Akihiro Kubo, Yuji Yasui, and Shin Ishii. Off-policy actor-critic for adversarial observation robustness: Virtual alternative training via symmetric policy evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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*, 2022.

Alexander Nikulin, Ilya Zisman, Denis Tarasov, Nikita Lyubaykin, Andrei Polubarov, Igor Kiselev, and Vladislav Kurenkov. Latent action learning requires supervision in the presence of distractors. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Simone Parisi, Alireza Kazemipour, and Michael Bowling. Beyond optimism: Exploration with partially observable rewards. In *Advances in Neural Information Processing Systems*, 2024.

Emilio Parisotto, H. Francis Song, Jack W. Rae, Razvan Pascanu, Caglar Gulcehre, et al. Stabilizing transformers for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Wei Wang, Dongqi Han, Xufang Luo, and Dongsheng Li. Addressing signal delay in deep reinforcement learning. In *International Conference on Learning Representations*, 2024.

Zeyu Wang, Yao-Hui Li, Xin Li, Hongyu Zang, Romain Laroche, and Riashat Islam. Learning fused state representations for control from multi-view observations. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Ziyuan Zhou, Guanjun Liu, Mengchu Zhou, and Weiran Guo. Robust multi-agent reinforcement learning with stochastic adversary. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

A Supplementary Experimental Specification

A.1 Diagnostic Task Factors

The controlled suite varies factors independently rather than binding relevance to a fixed feature identity. Corruption is normalized within each observation channel. A reference encoder–predictor trained on the development split is frozen before constructing the matched-surprise set, which is then reused for every method. Candidate states are stratified by task, corruption family/severity, cue-to-decision delay, token identity, and token-frequency bin before matching reference innovation. The reference innovation is used only to construct the evaluation set, while decision relevance comes from the controlled environment oracle in Eq. (12).

Table 5: Controlled diagnostic factor families and level counts.

Factor	Level count	Matched quantity	Changed quantity	Primary metric
Corruption	5	relevance, delay	perturbation strength	return
Relevance	5	surprise, token	action-value impact	$\Delta\alpha$
Delay	5	surprise, relevance	cue-to-decision gap	U-regret
Aliasing	4	observation norm	latent ambiguity	state probe
Distractor	4	signal SNR	nuisance identity	nuisance probe
Context	4	same token	decision consequence	AUROC

A.2 Training Hyperparameters

Table 6: Training hyperparameters for the recurrent PPO backbone and gate.

Parameter	Value	Parameter	Value	Parameter	Value
Discount γ	0.99	GAE λ	0.95	PPO clip	0.2
Belief dim.	256	Gate hidden	128	Learning rate	3×10^{-4}
Probe horizon H	8	Probe rate p_{probe}	0.05	Paired seeds M	2
$\lvert\mathcal{B}\rvert$	5	Revision penalty κ	0.01	Utility scale c	0.25
λ_g	1.0	λ_p	0.1	Min. weight ϵ	0.1
Target EMA	0.995	Final seeds	10	Dev seeds	5
η_{fail}	0.25	τ_α	dev-selected, frozen	Pair strata	fixed pre-final

A.3 Baseline Matching and Information Constraints

Table 7: Baseline matching rules and information constraints.

Method family	Capacity match	Training-only information	Evaluation constraint
Recurrent baselines	total params $\pm 5\%$	none	same observations
Memory modules	total params $\pm 5\%$	none	same history window
Privileged methods	closest matched width	latent/global state if required	no privileged test input
Robust methods	encoder + actor budget	perturbation generator if required	identical shift suite
BeliefGate	recurrent backbone + gate	paired simulator branches	single-pass test policy
Oracle gate	same deployment model	long-MC utility on 21-point grid	diagnostic only

A.4 Observation Interventions

Table 8: Observation interventions used for robustness and cross-shift tests.

Intervention	Training status	Operation	Matching rule	Failure mode tested
Cue swap	Seen	exchange task cue tokens	equal frequency	semantic relevance
Missing observation	Seen	replace selected frames by mask	equal missing rate	evidence absence
Burst noise	Seen	correlated additive corruption	equal average energy	temporal corruption
Aliasing	Held out	merge observation identities	equal marginal frequency	hidden-state ambiguity
Delay	Held out	postpone informative cue	equal cue content	long memory
Distractor	Seen	inject nuisance feature	equal distractor magnitude	irrelevance filtering

For vector-valued distractors, matched pairs preserve nuisance magnitude (e.g., $\|d_t^{(1)}\|_2 = \|d_t^{(2)}\|_2$) while changing nuisance identity or content.

A.5 Frozen-Representation Probes

All methods receive the same held-out replay prefixes before probe fitting; policies are frozen and probe data are balanced by latent state and intervention type.

Table 9: Frozen-representation probes used to audit belief content.

Probe target	Head	Metric	Interpretation
Discrete latent state	linear softmax	accuracy	state recoverability
Continuous latent state	linear ridge	R^2	metric state content
Next critical event	linear softmax	AUROC	decision-relevant memory
Distractor identity	linear softmax	accuracy	nuisance retention
Action at branch point	linear softmax	accuracy	control sufficiency

A.6 Sensitivity and Reporting Grid

Table 10: Sensitivity sweep and reporting statistics.

Quantity	Sweep	Held fixed	Report
Probe horizon H	$\{4, 8, 16\}$	probe rate	return, target agreement
Probe rate p_{probe}	$\{.02, .05, .10\}$	horizon	branch cost, U-regret
Revision penalty κ	$\{0, .005, .01, .02\}$	c	$\Delta\alpha$, return
Utility scale c	$\{.10, .25, .50, 1.0\}$	ϵ	gate-loss stability
Candidate grid	3, 5, 9 levels	compute budget	regret, cost
Prediction weight λ_p	$\{0, .05, .1, .2\}$	gate loss	return, state probe

A.7 Full Robustness Breakdown

Table 11: Per-intervention normalized return on the six base observation shifts.

Method	Clean	Cue swap	Missing	Burst	Alias	Delay	Distractor	Worst	Mean	IQM
GRU-PPO	0.756	0.618	0.566	0.528	0.491	0.458	0.604	0.458	0.544	0.551
R2D2-style	0.788	0.665	0.613	0.579	0.548	0.512	0.641	0.512	0.593	0.599
Memoroid	0.812	0.706	0.662	0.626	0.584	0.571	0.681	0.571	0.638	0.644
Stable Hadamard Memory	0.846	0.744	0.701	0.667	0.632	0.655	0.712	0.632	0.685	0.691
Memory Traces	0.835	0.721	0.681	0.648	0.603	0.692	0.673	0.603	0.670	0.676
Policy-POMDP (2025)	0.824	0.736	0.653	0.621	0.688	0.646	0.669	0.621	0.669	0.676
Adv.-Obs. AC	0.782	0.704	0.726	0.738	0.602	0.553	0.749	0.553	0.679	0.687
BeliefGate	0.906	0.844	0.819	0.796	0.768	0.781	0.835	0.768	0.807	0.814

A.8 Detailed Computational Profile

Table 12: Detailed training and deployment cost profile for matched recurrent and gate baselines.

Method	Params (M)	FLOPs/step	Branch ratio	Extra sim. steps	Peak mem.	Train time	Train overhead	Infer. ms	Test branches
GRU-PPO	1.52	0.41G	0.00	0	1.80 GB	3.6 h	1.00×	0.31	0
R2D2-style	1.55	0.48G	0.00	0	2.04 GB	4.3 h	1.19×	0.36	0
Memoroid	1.49	0.45G	0.00	0	1.92 GB	4.0 h	1.11×	0.34	0
Stable Hadamard Memory	1.54	0.53G	0.00	0	2.21 GB	4.7 h	1.31×	0.38	0
Memory Traces	1.47	0.39G	0.00	0	1.73 GB	3.9 h	1.08×	0.29	0
PPO-only gate	1.50	0.46G	0.00	0	1.94 GB	4.1 h	1.14×	0.33	0
BeliefGate	1.53	0.50G	4.00	4.0M	2.48 GB	8.5 h	2.36×	0.35	0

A.9 Per-Task Benchmark Results

The main paper reports task-family IQM; Table 13 decomposes the aggregate over representative tasks from all three benchmark groups. BeliefGate has the highest representative-task IQM and is stronger on CueSwitch, AliasChain, KeyDoor, and MemoryNav, while Stable Hadamard Memory is stronger on RepeatFirst and CountRecall and Memory Traces on DelayedCue. The task-level results preserve the memory-specific strengths seen in the aggregate comparison.

Table 13: Representative per-task normalized return across all three benchmark groups.

Method	CueSwitch	AliasChain	KeyDoor	RepeatFirst	CountRecall	DelayedCue	MemoryNav	IQM
GRU-PPO	0.704	0.598	0.662	0.612	0.548	0.526	0.571	0.603
R2D2-style	0.731	0.681	0.707	0.654	0.623	0.608	0.646	0.662
Memoroid	0.762	0.694	0.748	0.781	0.739	0.668	0.701	0.731
Stable Hadamard Memory	0.793	0.742	0.776	0.842	0.816	0.736	0.764	0.787
Memory Traces	0.754	0.706	0.731	0.802	0.773	0.861	0.808	0.789
Policy-POMDP (2025)	0.786	0.768	0.759	0.724	0.691	0.741	0.749	0.752
Adv.-Obs. AC	0.716	0.637	0.681	0.624	0.586	0.561	0.603	0.632
BeliefGate	0.914	0.861	0.892	0.831	0.806	0.852	0.824	0.858

A.10 Representation-Content Results

All probe heads are fit on the same frozen replay prefixes. Higher distractor-identity accuracy indicates greater nuisance retention, whereas the remaining metrics measure control-relevant state content. BeliefGate attains the highest critical-event AUROC and branch-action accuracy and the lowest distractor-identity accuracy, while memory-focused baselines remain stronger on generic latent-state accuracy or R^2. The pattern therefore concentrates the representation gains on decision-relevant content and nuisance suppression rather than uniformly richer state reconstruction.

Table 14: Frozen-representation probe results on a common held-out prefix set.

Method	Latent acc.↑	Latent R^2 ↑	Critical-event AUROC↑	Distractor acc.↓	Branch-action acc.↑
GRU-PPO	0.712	0.603	0.671	0.612	0.702
R2D2-style	0.746	0.641	0.729	0.584	0.756
Memoroid	0.789	0.682	0.774	0.552	0.721
Stable Hadamard Memory	0.856	0.748	0.821	0.526	0.768
Memory Traces	0.818	0.791	0.842	0.571	0.846
Policy-POMDP (2025)	0.838	0.726	0.808	0.497	0.872
Adv.-Obs. AC	0.698	0.657	0.748	0.438	0.739
BeliefGate	0.832	0.774	0.916	0.312	0.901

B Binary-Target Special Case and Utility Geometry

For $\mathcal{B} = \{0, 1\}$, the target selected by penalized utility in Eq. (6) becomes a hard comparison between preserving the prior and accepting the proposal. The deployed gate remains continuous in $[0, 1]$; only the supervision candidate set is binary. A soft-mean alternative would be equivalent to

$$\alpha_t^{\text{soft}} = \sigma\left(\frac{\hat{U}_t(1) - \hat{U}_t(0) - \kappa}{\tau}\right),$$

which assumes sufficiently regular utility between the endpoints. The selected-target formulation avoids this regularity assumption. The 21-point audit reports (i) the fraction of probe states with one local maximum, (ii) the average number of local maxima, and (iii) regret of $\beta_t^\star$ against the dense MC best. The oracle utility gate selects the maximizing revision strength on the 21-point grid using long-MC utility and is used only as a diagnostic upper bound.

C Additional Related Work Context

Several adjacent lines of work motivate decision-relevant state updates. Bisimulation metrics formalize control-relevant similarity (Ferns et al., 2011); partially observed reward exploration separates missing information from generic uncertainty (Parisi et al., 2024); and latent-action supervision shows how nuisance structure can confound representation learning (Nikulin et al., 2025). Robust and adversarial methods study counterfactual or stochastic perturbations (Belaire et al., 2025; Zhou et al., 2025), while multi-view and multi-agent methods add complementary information channels (Wang et al., 2025; Kang et al., 2025). BeliefGate focuses on how strongly the current observation should revise the learned belief representation used for the next decision.